Gestures in Human-Robot Interaction: Development of Intuitive Gesture Vocabularies and Robust Gesture Recognition

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium (Dr. rer. nat.)
im Fach Informatik

eingereicht an der

Mathematisch-Naturwissenschaftliche Fakultät
der Humboldt-Universität zu Berlin

von M.Sc. **Saša Bodiroža**

Präsident der Humboldt-Universität zu Berlin:

Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:

Prof. Dr. Elmar Kulke

Gutachter(innen):

1. Prof. Dr. Verena V. Hafner	Humboldt-Universität zu Berlin
2. Prof. Dr. Yael Edan	Ben-Gurion University of the Negev (Israel)
3. Prof. Dr. Bruno Lara	Universidad Autónoma del Estado de Morelos (Mexico)

Tag der Verteidigung: 17.03.2016

Abstract

Gestures consist of movements of body parts and are a mean of communication that conveys information or intentions to an observer. Therefore, they can be effectively used in human-robot interaction, or in general in human-machine interaction, as a way for a robot or a machine to infer a meaning, e.g., a person's particular intention or state. In order for people to intuitively use gestures and understand robot gestures, it is necessary to define mappings between gestures and their associated meanings – a gesture vocabulary. The task is to determine human and robot gesture vocabularies. In the former case, the vocabulary defines which gestures a group of people would intuitively use to convey information, while in the latter case, it displays which robot gestures are deemed as fitting for a particular meaning. Once these vocabularies are defined, their effective use depends on techniques for gesture recognition, which considers classification of body motion into discrete gesture classes, relying on pattern recognition and machine learning.

This thesis addresses both research areas, presenting development of gesture vocabularies as well as gesture recognition techniques, focusing on hand and arm gestures. Prior to the gesture vocabularies and recognition, attentional models for humanoid robots were developed as a prerequisite for human-robot interaction and a precursor to gesture recognition. A method for defining gesture vocabularies for humans and robots, based on user observations and surveys, is explained and experimental results are presented. As a result of the robot gesture vocabulary experiment, an evolutionary-based approach for refinement of robot gestures is introduced, based on interactive genetic algorithms. A robust and well-performing gesture recognition algorithm based on dynamic time warping has been developed. Most importantly, it employs one-shot learning, meaning that it can be trained using a low number of training samples and employed in real-life scenarios, lowering the effect of environmental constraints such as relative positions and orientations of the sensor and the person, and gesture features. Finally, an approach for learning a relation between self-motion and pointing gestures is presented.

Zusammenfassung

Gesten sind Bewegungen von Körperteilen, und ein Kommunikationsweg, der einem Betrachter Informationen oder Absichten übermittelt. Daher können sie effektiv in der Mensch-Roboter-Interaktion, oder in der Mensch-Maschine-Interaktion allgemein, verwendet werden. Sie stellen eine Möglichkeit für einen Roboter oder eine Maschine dar, um eine Bedeutung abzuleiten, zum Beispiel den Zustand oder die Absicht einer Person. Um Gesten intuitiv benutzen zukönnen und Gesten, die von Robotern ausgeführt werden, zu verstehen, ist es notwendig, Zuordnungen zwischen Gesten und den damit verbundenen Bedeutungen zu definieren – ein Gestenvokabular. Es geht darum, Mensch- und Robotergestenvokabulare festzulegen. Im ersten Fall definiert das Vokabular welche Gesten ein Personenkreis intuitiv verwendet, um Informationen zu übermitteln. Im zweiten Fall zeigt es, welche Robotergesten zu welcher Bedeutung passen. Sind diese Vokabulare einmal definiert, hängt ihre effektive und intuitive Benutzung von Gestenerkennung ab, das heißt von der Klassifizierung der Körperbewegung in diskrete Gestenklassen durch die Verwendung von Mustererkennung und maschinellem Lernen.

Die vorliegende Dissertation befasst sich mit beiden Forschungsbereichen, das heißt mit der Entwicklung von Gestenvokabularen sowie Gestenerkennungstechniken, wobei der Fokus auf Hand- und Armgesten liegt. Als eine Voraussetzung für die intuitive Mensch-Roboter-Interaktion wird zunächst ein Aufmerksamkeitsmodell für humanoide Roboter entwickelt. Danach wird ein Verfahren für die Festlegung von Mensch- und Roboter-Gestenvokabulare vorgelegt, das auf Beobachtungen von Benutzern und Umfragen beruht. Anschliessend werden experimentelle Ergebnisse vorgestellt. Eine Methode zur Verfeinerung der Robotergesten wird entwickelt, die auf interaktiven genetischen Algorithmen basiert. Ein robuster und performanter Gestenerkennungsalgorithmus wird entwickelt, der auf Dynamic Time Warping (dynamischer Zeitverzerrung) basiert, und sich durch die Verwendung von One-Shot-Learning auszeichnet, das heißt durch die Verwendung einer geringen Anzahl von Trainingsgesten. Der Algorithmus kann in realen Szenarien verwendet werden, womit er den Einfluss von Umweltbedingungen, zum Beispiel die relativen Positionen und Orientierungen des Sensors und der Person, sowie Gesteneigenschaften, senkt. Schließlich wird eine Methode für das Lernen der Beziehungen zwischen Selbstbewegung und Zeigegesten vorgestellt.

Acknowledgements

The work towards this dissertation was an interesting, enjoyable and at times challenging journey. I am thankful for the support of the people who were around me throughout my doctoral studies at the Humboldt-Universität zu Berlin. Most of all, I would like to express my sincere gratitude to Prof. Verena V. Hafner, who provided me with guidance, as well as with freedom in my research. Her interdisciplinary approach to human-robot interaction inspired me and shaped this thesis in large part. Furthermore, I am thankful to Prof. Yael Edan, Prof. Helman Stern and Prof. Bruno Lara for their support, numerous productive discussions and helpful comments. I am in particular thankful to Prof. Yael Edan for her hospitality during my research visit at the Ben-Gurion University of the Negev.

Working at Verena's group was an insightful, inspiring and creative experience. Sharing the office and working together with Guido Schillaci has made the days in Adlershof so much more enjoyable. Furthermore, I had the pleasure to meet, work and spend nice time with researchers and students in the group – Siham Al-Rikabi, Ferry Bachmann, Oswald Berthold, Christian Blum, Damien Drix, Lovísa Helgadottír, Ivana Kajić, Heinrich Mellmann, Aleke Nolte, Antonio Pico Villalpando, Claas Ritter, Marcus Scheunemann, Benjamin Schlotter. Renate Zirkelbach made the university bureaucracy a navigable and pleasant environment and I have to thank her for all interesting chats.

A half-year research visit at Yael's group at the Ben-Gurion University provided a fresh influence on my research, in particular related to the design of gesture vocabularies. I am fully grateful to her and Helman for their support and pleasant and constructive discussions. The stay in Israel was only made better by wonderful people I have met there – in particular to Guillaume Doisy, Pauline Marchand, Yael Ron, Tamir David Hod, Tanya Levi, Danit Nativ, Maayan Sivan.

Being part of the INTRO project, an international research network, enabled me to look on research topics in HRI from various aspects and I would like to thank fantastic project colleagues – Guido Schillaci, Guillaume Doisy, Aleksandar Jevtić, Maria Elena Giannaccini, Bo Li, Benjamin Fonooni and Roy Someshwar. I'm grateful I had a chance to meet you and to work with you.

During the time spent in Berlin, I began to feel here as in my second hometown. It probably wouldn't be like that if it wasn't for friends I made here, without whom the time spent in this city would not be nearly as fun. I would like to thank in particular to Rajnish Rao, Ziv Harpaz, Ivana Jerković, Evgeny Bobrov, Mirko Tadić, Bratislav Milić, Smadar Ovadia-Caro, Luis Miranda.

I was always looking forward to going back to Belgrade and spending time with my friends – Jovana Martinović, Jovana Dačković, Ivan Perkov, Minja Davidović, Jovana Avalić, Aleksandar Mirić, Goran Petrović, Mihajlo Stojković, Miloš Đerić, Jelena Ivanov, Milena Naumović, Blažo Bošković, Filip Maljković. Thank you for being amazing and for bearing with me, for all fun moments we had and for all those that will come.

My parents Milan and Milosava, and my sister Sanja have provided tremendous support. They have always been full of patience, as well as supporting and kind words and I cannot express how thankful I am to them for everything they provided for me. Finally, I would not be who I am today without my partner Ivan. Thank you for being there and for being my best friend.

The research leading to these results was supported by the European Commission's Seventh Framework Programme under the INTRO ITN project (grant agreement no. 238486) and under the EARS project (grant agreement no. 609465), by the German Research Foundation within the GRK 1589/1 "Sensory Computation in Neural Systems", and by the Paul Ivanier Center for Robotics Research and Production Management, and the Rabbi W. Gunther Plaut Chair in Manufacturing Engineering, Ben-Gurion University.

Contents

List of Figures iii

List of Tables v

1 Introduction 1
1.1 Background . . . 1
1.2 Objectives . . . 2
1.3 Methodology . . . 2
1.4 Contribution . . . 4
1.5 Structure . . . 5
1.6 Publications . . . 5

2 Defining Gestures 8
2.1 Gesture Taxonomies in Human-Robot Interaction and Human-Computer Interaction 11

3 Attention Systems for Human-Robot Interaction 14
3.1 Overview . . . 14
3.2 Background . . . 15
3.3 Implementation of an Attentional Mechanism for a Humanoid Robot . . . 17
3.3.1 Saliency Detection and Attention Manipulation . . . 17
3.3.2 Behaviors . . . 22
3.3.3 Robot Platform . . . 23
3.3.4 Experimental Setup . . . 24
3.4 Discussion and Conclusions . . . 31

4 Gesture Vocabularies for Human-Robot Interaction 38
4.1 Overview . . . 38
4.2 Background . . . 38
4.3 Defining Gesture Vocabularies . . . 41
4.4 Case Study: Gesture Vocabularies for a Robot Waiter Scenario . . . 42
4.4.1 Human Gesture Vocabulary . . . 43
4.4.2 Robot Gesture Vocabulary . . . 46
4.4.3 Comparison of Present Actions in Human and Robot Gesture Vocabularies 58
4.5 An Evolutionary Approach for Improving Robot Gestures . . . 58
4.5.1 Genetic Algorithms . . . 58
4.5.2 Interactive Genetic Algorithms . . . 60
4.5.3 Evolution of Robot Gestures . . . 60
4.6 Discussion and Conclusions . . . 65
4.6.1 Development of Gesture Vocabularies . . . 65

4.6.2 Evolution of Robot Gestures . . . 66

5 Gesture Recognition and Disambiguation 67
5.1 Overview . . . 67
5.2 Background . . . 69
5.3 Recognition, Representation, Application and Advantages . . . 69
5.4 Framework for Gesture Recognition and Disambiguation . . . 71
5.4.1 Data Acquisition . . . 71
5.4.2 Data Preprocessing . . . 72
5.4.3 Gesture Recognition . . . 74
5.4.4 Gesture Disambiguation . . . 78
5.5 Performance Evaluation . . . 79
5.5.1 Evaluation of the Algorithm . . . 79
5.6 Gesture-based Control of a Person-following Robot . . . 83
5.6.1 Background . . . 83
5.6.2 Framework for Extended Human-Robot Interaction . . . 84
5.6.3 Person Following by a Robot Controlled with Gestures . . . 85
5.7 Internal Models for Gesture Recognition . . . 85
5.7.1 Inverse and Forward Models for Action Execution . . . 86
5.7.2 Learning Motor Control for Rotation Based on Arm Movement . . . 87
5.7.3 Learning Motor Control for Rotation and Translation Based on Pointing Gestures . . . 89
5.8 Discussion and Conclusions . . . 91

6 Conclusions and Future Work 98
6.1 Attentional Models . . . 98
6.2 Gesture Vocabularies . . . 98
6.3 Gesture Recognition and Internal Models . . . 99
6.4 Future Work . . . 100
6.4.1 Attentional Models . . . 100
6.4.2 Gesture Vocabularies . . . 100
6.4.3 Gesture Recognition and Internal Models . . . 101

A Questionnaires 102
A.1 Perception of Robot Behavior . . . 102
A.1.1 Questionnaire . . . 102
A.2 Human Gesture Vocabulary . . . 104
A.2.1 Consent Form . . . 104
A.3 Robot Gesture Vocabulary . . . 104
A.3.1 Consent Form . . . 104
A.3.2 Questionnaire . . . 105

Bibliography 106

List of Figures

1.1 Flow diagram of the procedure for using gestures in human-robot interaction . . . 3

2.1 Gesture levels . . . 9
2.2 Kendon's continuum . . . 10
2.3 Gesture spaces . . . 11
2.4 Frequency of gestures according to their type . . . 12

3.1 Overview of the attentive mechanism. . . . 18
3.2 Illustration of an ego-sphere representation. . . . 19
3.3 A typical babbling sequence using the Nao platform. . . . 21
3.4 Example of a sequence of pointing gestures performed by the robot. . . . 21
3.5 Nao with colored stickers that indicate the position of the fake eyes. . . . 23
3.6 Experimental setup showing interaction between the Nao and a person. . . . 25
3.7 Results of the experiment from the Godspeed questionnaire . . . 29

4.1 Research topics covered in Chapter 4 . . . 39
4.2 Example of a designed gesture vocabulary for a modeled world, and a gesture vocabulary seen in the real world . . . 42
4.3 Procedure for obtaining human and robot gesture vocabularies . . . 43
4.4 Obtained human gesture vocabulary, with excluded gesture 11 ("no gesture") and meaning 11 ("Where is the toilet?") . . . 46
4.5 Rankings and mean overall impressions of gesture alternatives for action 1 . . . 52
4.6 Rankings and mean overall impressions of gesture alternatives for action 2 . . . 53
4.7 Rankings and mean overall impressions of gesture alternatives for action 3 . . . 53
4.8 Rankings and mean overall impressions of gesture alternatives for action 4 . . . 54
4.9 Rankings and mean overall impressions of gesture alternatives for action 5 . . . 55
4.10 Rankings and mean overall impressions of gesture alternatives for action 6 . . . 55
4.11 Rankings and mean overall impressions of gesture alternatives for action 7 . . . 56
4.12 Rankings and mean overall impressions of gesture alternatives for action 8 . . . 56
4.13 Experimental environment for evolution of robot gestures . . . 62
4.14 A decrease in mean fitness value for Action 7 during one experimental run. . . . 64
4.15 An example of the trend of mean fitness value for a successful evolution. . . . 64
4.16 An example of the trend of mean fitness value for an unsuccessful evolution. . . . 65

5.1 Research topics covered in Chapter 5 . . . 68
5.2 Proposed Gesture Recognition and Disambiguation (*GRaD*) framework . . . 72
5.3 Illustration of direction invariance . . . 75
5.4 Robot platform robuLAB10. . . . 86

5.5 Example of a path-following algorithm. 86
5.6 Internal models. 87
5.7 Representation of the learning process 89
5.8 Representation of the execution process 90
5.9 Illustration of the experimental setup for motor control learning. 91
5.10 Experimental setup for motor control learning. 92
5.11 Illustration of the interconnected self-organizing maps. 92
5.12 Weights of input dimensions of self-organizing maps encoding perceived pointing (SOM 1) and related motor commands (SOM 2) 93

List of Tables

3.1 Most relevant correlations (part 1). 34
3.2 Most relevant correlations (part 2). 35
3.3 Statistically significant results of RM ANOVA on the questionnaire variables. . . . 36
3.4 Statistically significant results of RM ANOVA on the proxemics variables. 37

4.1 Results of the human gesture vocabulary experiment. 45
4.2 Selected actions and their representative gestures 48
4.3 Correlations for gesture alternatives of action 1. 49
4.4 Correlations for gesture alternatives of action 2. 49
4.5 Correlations for gesture alternatives of action 3. 50
4.6 Correlations for gesture alternatives of action 4. 50
4.7 Correlations for gesture alternatives of action 5. 50
4.8 Correlations for gesture alternatives of action 6. 51
4.9 Correlations for gesture alternatives of action 7. 51
4.10 Correlations for gesture alternatives of action 8. 51
4.11 Means and standard deviations for overall impression, precision and speed of gesture alternatives of action 1. 52
4.12 Means and standard deviations for overall impression, precision and speed of gesture alternatives of action 2. 53
4.13 Means and standard deviations for overall impression, precision and speed of gesture alternatives of action 3. 54
4.14 Means and standard deviations for overall impression, precision and speed of gesture alternatives of action 4. 54
4.15 Means and standard deviations for overall impression, precision and speed of gesture alternatives of action 5. 54
4.16 Means and standard deviations for overall impression, precision and speed of gesture alternatives of action 6. 55
4.17 Means and standard deviations for overall impression, precision and speed of gesture alternatives of action 7. 56
4.18 Means and standard deviations for overall impression, precision and speed of gesture alternatives of action 8. 57
4.19 Resulting robot gesture vocabulary 57
4.20 Results of the first experiment on gesture evolution with first success criterion . . . 62
4.21 Results of the first experiment on gesture evolution with second success criterion . 63
4.22 Results of the first experiment on gesture evolution with third success criterion . . 63
4.23 Results of the verification experiment on gesture evolution 63

5.1 Issues in gesture recognition 71

5.2 Confusion matrix of the gesture recognition without preprocessing. 80
5.3 Confusion matrix of the gesture recognition with frame of reference transformation. 81
5.4 Confusion matrix of the gesture recognition with frame of reference transformation and alignment. 81
5.5 Confusion matrix of the gesture recognition with frame of reference transformation and scaling. 81
5.6 Confusion matrix of the gesture recognition with frame of reference transformation, alignment and scaling. 82
5.7 Summary of evaluation results 82
5.8 Confusion matrix of the gesture recognition from the Sehir University gesture dataset (results from paper). 83
5.9 Confusion matrix of the gesture recognition from the Sehir University gesture dataset. 83
5.10 Fixed PT-M with no displacement 94
5.11 Fixed PT-M with displacement of $0.5m$ to the left 94
5.12 Fixed PT-M with displacement of $0.5m$ to the right 95
5.13 Fixed PT-M with displacement of $1m$ to the left 95
5.14 Fixed PT-M with displacement of $1m$ to the right 95
5.15 Moving PT-M, distance of $2m$ from the robot 96
5.16 Robot moving forward toward the person 96
5.17 Robot moving perpendicular to the person 96
5.18 Robot person-following control 96
5.19 Average error during testing of the motor control learning. 97

Chapter 1

Introduction

1.1 Background

Humans rely on various communication methods, such as speech and gestures. Traditionally, communication channels in human-machine interaction do not fully overlap with those used in interpersonal communication, which can result in decrease of system's intuitiveness and increase in perceived workload during the interaction. Recently, a shift towards more natural and intuitive ways of interaction is observed, such as in use of natural language processing and speech synthesis for issuing commands and providing feedback, respectively.

Another intuitive form of communication are gestures. A gesture is defined as a movement of body parts, most commonly hands and head, with the goal of communicating a particular message (Mitra and Acharya, 2007). They are observed in early development of infants, such as imperative and declarative pointing to specify a particularly interesting point in their environment, prior to development of speech. Furthermore, there is evidence for gesturing in other primates (Tomasello et al., 1997; Pika et al., 2003; Liebal et al., 2006). Due to their ubiquitous use in interpersonal interaction, they could be employed as a natural and intuitive way in human-machine, and in particular in human-robot interaction.

A different form of gestures is already used in human-computer interaction, where a gesture represents motions performed with fingers on touch-sensitive surfaces. However, the focus of this thesis is on use of gestures in human-robot interaction, focusing on the use of gestures that occur naturally in interpersonal interaction. Use of gestures in human-robot interaction is investigated through three, inter-connected areas.

The first area considers methods of developing gesture vocabularies, mappings between gestures and their associated meanings . There are human and robot gesture vocabularies, and partial overlap between those can appear. Gesture vocabularies depend on use-case scenario and features of the people that will interact with the robot, as well as on a morphology of the robot, which defines which gestures can be performed and correctly recognized by a human observer.

The second area deals with gesture recognition. Gesture recognition is a research area within machine learning and pattern recognition, covering detection, segmentation and recognition of gestures in human motion. Its main application is in human-computer, and later in human-robot interaction, as a way to provide an intuitive interface. In human-computer interaction, communication with gestures was unidirectional, as computers do not posses required capacities to gesture. On the other side, some robots posses limbs which can be used to generate gestures, in order to convey a message in a more intuitive way to a human listener. This poses an open question of defining which gestures should the robot use and how should they be designed. As a result,

there is a need for a comprehensive study on the use of gestures in human-robot interaction, taking in consideration various gesture-related aspects, with the goal of having an intuitive interaction, which defines the main topic of this thesis.

The third area is gesture synthesis, which comprises of methods of how and when should a robot produce a gesture. It includes generative models controlling when a robot gestures, which can be related to robot's particular state, or what is it currently saying. However, gesture synthesis is partially out of the scope of this work.

1.2 Objectives

Having in mind above outlined areas in gesture use in human-robot interaction, there are three main research questions that are addressed by the thesis:

- How to develop gesture vocabularies, that is sets of gestures mapped to their associated meanings for a particular interaction scenario?
- How can gestures be learned, recognized and reproduced?
- How can gestures be disambiguated?

Having in mind that gestures differ between different groups of people, effective methods are needed for development of mappings between gestures and actions they represent, so called gesture vocabularies. Identified gestures need to be analyzed to see whether different gesture classes from a vocabulary can be distinguished and successfully recognized.

Following from the previous, the aims of the thesis are:

- Study gestures that occur in interpersonal interactions and introduce an effective method for development of gesture vocabularies,
- Develop a system for gesture recognition and disambiguation,
  - Develop an algorithm for dynamic gesture recognition,
  - Propose an approach for gesture disambiguation, which can follow the recognition phase.

1.3 Methodology

Use of gestures in human-robot interaction requires a multifaceted approach, as represented with the flow diagram in Figure 1.1. While some gestures will be used across different scenarios, some will be specific for a particular scenario. Therefore, the first part of the process is to define a use-case scenario. For example, the use-case scenario used here is the robot waiter scenario, which describes the interaction occurring between a customer and a waiter in a bar or a restaurant. A set of actions, both user-initiated, such as asking for a menu, and robot-initiated, such as offering a suggestion, can be extracted from the scenario. The gestures coupled with these actions form a gesture vocabulary.

There is more than one approach to selecting gestures to represent the actions. A simple way would be to have a system developer to design gestures and predefine the mappings between gestures and actions. However, with this approach the user is taken out of the loop. A better way would be to look at recorded instances of extracted actions and use the gestures that occur within the recordings. Alternatively, users could be asked to rank and select gestures from a defined set of gestures, or a sample of the user group could be surveyed to find out which gestures they

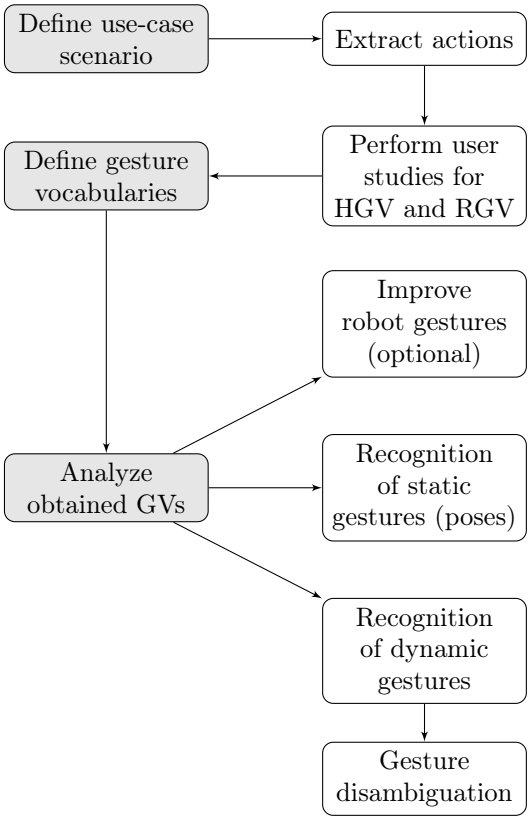


Figure 1.1: Flow diagram of the procedure for using gestures in human-robot interaction

associate with the actions. The latter approach is used in this work. Firstly, it directly interacts with persons who would interact with the robot. This direct approach might provide gestures that better represent gestures that would be used. A disadvantage of this approach is that it is probable that the resulting mapping between gestures and actions will be many-to-many. However, depending on the particular case, this might not be an issue, if surrounding context is taken into account.

Analysis of the results of the user studies performed for development of human and robot gesture vocabularies drives the further part of the process. In case of gestures appearing in the robot gesture vocabulary (RGV), the results of the study might not be satisfactory – the participants might not see the gestures as good enough. In this case an interactive procedure can be applied for improvement of robot gestures. Interactive genetic algorithm is an appropriate choice, as it can be used in cases where aesthetic qualities are being used as a rating. Gestures appearing in human gesture vocabulary (HGV) drive the development of the gesture recognition and disambiguation systems. In some cases, the vocabulary will contain static gestures, or poses. However, the focus in this work is on the recognition of dynamic gestures.

There are prior methods for recognition of dynamic gestures. However, there was lack of focus on application of gesture recognition approaches in real-life cases. Some of the issues were relying on large training datasets or ignoring some of the environmental issues. Therefore, the approach taken here is based on one-shot learning, where learning algorithms rely on small training datasets. Furthermore, in order to alleviate some of the environmental factors, a pre-processing pipeline was envisioned to improve correct recognition under different situations. The disadvantage of one-shot learning approaches is that they are highly dependent on correct performance of training samples. To evaluate the gesture recognition algorithm, training and test data was collected.

Some other areas are studied in this thesis. Namely, the effect of attention models on perception of the robot was studied in a user survey. This is identified as one of the prerequisites for natural human-robot interaction. Godspeed questionnaire was selected as a standardized method for evaluation of robot's appearance and behavior. This can show how particular robot's behavior affect how the others perceive it. An interesting question arose during the research of gesture recognition. Namely, it would be interesting to see how can relation between gestures and actions be learned. Two experiments were performed to see how can gestures lead to action execution.

1.4 Contribution

The main contributions of the work presented in the thesis are:

- An approach for development of human and robot gesture vocabularies based on user studies, thus increasing the intuitiveness of gestures.
  - Human and robot gesture vocabularies were developed for an example robot waiter scenario.
- An approach based on interactive genetic algorithms for improvement of robot gestures according to particular aesthetic preferences of a user or a group of users.
- System for gesture recognition and disambiguation focused on real-life use.
  - One-shot learning algorithm for gesture recognition requires very low number of gesture samples for training and is highly robust to environmental interferences, such as position and orientation of the person relative to the robot, gesture location, size of the person, and therefore the size of the gesture, and the speed of performed gestures.

1.5 Structure

The rest of the thesis is organized as follows. Chapter 2 presents a general literal review on what gestures are and how they are defined in literature, as well as present gesture taxonomies in gesture studies. Furthermore, the concept of a gesture space, in which the gesturing occurs, is explained, showing that gestures are not limited to a particular location in gesture space across different persons. Chapter 3 introduces the notion of attentional models and presents an implementation of an attentional model for a humanoid robot, as a prerequisite for intuitive human-robot interaction. The effect of different behaviors on the way others perceive the robot was tested to identify the influence of particular characteristics of robot's behaviors. Chapter 4 introduces the notion of gesture vocabularies and how they can be developed for both humans and robots. The procedure is applied on development of a human and a robot gesture vocabulary for a robot-waiter interaction scenario. Chapter 5 presents the gesture recognition and disambiguation framework. Firstly, development of a robust gesture recognition algorithm is presented, focusing on its use in real world scenarios. Most importantly, the algorithm applies one-shot learning, which means that the training is performed using low number of training samples. This is followed by presentation of a theoretical concept for gesture disambiguation. The second part of the chapter introduces an approach for learning the relation between visual sensory input and motor commands, using the example of pointing gestures. Finally, Chapter 6 discusses the obtained results, highlights main accomplishments, what are the important outcomes and what could be possible future research directions based on the presented results.

Due to the diverse research directions, chapters 3, 4 and 5 begin with overview sections, outlining what is presented in the chapter, followed by background sections, which cover the literature reviews in their respective fields, instead of having a separate chapter devoted to the literature review.

1.6 Publications

Following publications resulted from the work presented in this thesis.

Journal articles

Related to Chapter 3:

- Schillaci, G., **Bodiroža**, S. and Hafner, V. V. (2012), Evaluating the Effect of Saliency Detection and Attention Manipulation in Human-Robot Interaction, International Journal of Social Robotics, DOI: 10.1007/s12369-012-0174-7.
  All authors actively participated in the design and implementation of the experiments. Pointing algorithm was developed by Guido Schillaci.

Peer-reviewed conference proceedings

Related to Chapter 3:

- **Bodiroza**, S., Schillaci, G. and Hafner, V.V. (2011), Robot Ego-sphere: An Approach for Saliency Detection and Attention Manipulation in Humanoid Robots for Intuitive Interaction, in Proceedings of the 11th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2011), Bled, Slovenia.

Related to Chapter 4:

- **Bodiroža**, S., Stern, H. I., Edan Y. (2012), Dynamic Gesture Vocabulary Design for Intuitive Human-Robot Dialog, in Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2012), Boston, USA.

Related to Chapter 5:

- **Bodiroža**, S., Hafner, V. V. (2014), GRaD: Gesture Recognition and Disambiguation Framework for Unconstrained, Real-Life Scenarios, in Workshop Proceedings of 13th International Conference on Intelligent Autonomous Systems (IAS-13), pp. 347-353, Padua, Italy.

- Jevtić, A., Doisy, G., **Bodiroža**, S., Edan, Y., and Hafner, V. V. (2014), Human-Robot Interaction through 3D Vision and Force Control, in Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2014), p. 102, Bielefeld, Germany.
  Guillaume Doisy and Aleksandar Jevtić worked on person following and pointing algorithms. Aleksandar Jevtić integrated the separate parts into the final demo.

- **Bodiroža**, S., Jevtić, A., Lara, B. and Hafner, V. V. (2013), Learning the Relation of Motion Control and Gestures Through Self-Exploration, in Proceedings of the Robotics Challenges and Vision Workshop, Robotics: Science and Systems Conference (RSS 2013), Berlin, Germany.

- **Bodiroža**, S., Doisy, G. and Hafner, V. V. (2013), Position-Invariant, Real-Time Gesture Recognition Based on Dynamic Time Warping, in Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2013), Tokyo, Japan.
  Guillaume Doisy provided support with the transformation matrices.

- Doisy, G., Jevtić, A. and **Bodiroža**, S. (2013), Spatially Unconstrained, Gesture-Based Human-Robot Interaction, in Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2013), Tokyo, Japan.
  Guillaume Doisy and Aleksandar Jevtić worked on person tracking, position estimation and following algorithms.

Peer-reviewed conference presentations
Related to Chapter 4:

- **Bodiroža**, S., Stern, H. I., Hafner, V. V., Edan, Y. (2012), A Diachronic Approach for Human-Humanoid Discourse, presented at the 5th Conference of International Society for Gesture Studies, Lund, Sweden. abstract

Related to Chapter 5:

- **Bodiroža**, S., Jevtić, A., Lara, B. and Hafner, V. V. (2013), Learning Motion Control for Guiding a Robot using Gestures, presented at the 4th Israeli Conference on Robotics, Tel Aviv, Israel.

- Doisy, G., Jevtić, A., **Bodiroža**, S. and Edan, Y. (2013), Spatially Unconstrained Natural Interface for Controlling a Mobile Robot, presented at the 4th Israeli Conference on Robotics, Tel Aviv, Israel.
  Guillaume Doisy and Aleksandar Jevtić worked on person tracking, position estimation and following algorithms.

Following publications were published based on the results not related to the work presented in the thesis:

Peer-reviewed conference proceedings

- Kajić, I., Schillaci, G., **Bodiroža, S.** and Hafner, V. V. (2014), Learning Hand-Eye Coordination for a Humanoid Robot Using SOMs, in Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2014), pp. 192-193, Bielefeld, Germany.

- Kajić, I., Schillaci, G., **Bodiroža, S.** and Hafner, V.V. (2014), A Biologically Inspired Model for Coding Sensorimotor Experience Leading to the Development of Pointing Behaviour in a Humanoid Robot, Workshop on "HRI: a bridge between Robotics and Neuroscience" at HRI-2014, Bielefeld, Germany.

Code and anonymous datasets resulting from the work on this thesis are available upon request.

Chapter 2

Defining Gestures

In the scope of human-machine interaction, Mitra and Acharya (2007) define gestures as "expressive, meaningful body motions involving physical movements of the fingers, hands, arms, head, face, or body with the intent of conveying meaningful information, or interacting with the environment." Due to their impact in the communication, the work presented here focuses on arm gestures with intent of information transfer.

A typical gesture taxonomy would include *manipulative* and *communicative* gestures, where the latter can be further divided into *symbols*, which can be *referential* or *modalizing*, and *acts*, which can be *mimetic* or *deictic* (Quek, 1995). Communicative gestures are of higher interest with regards to the aspects of human-robot and human-computer interaction, and are therefore the main target for recognition and disambiguation.

From communicative stance, gestures can be observed within a narrower frame. Kendon (1986) defined a gesture as a movement of body, as already stated above, but with a clear goal of communicating a thought or emotional state. He specifies that gestures are *autonomous*, which can appear without speech, or *gesticulation*, appearing together with speech. McNeill (1986) defines a gesture as "any spontaneous body movement during speech (usually of the hands) not performed for a practical purpose, i.e., not a manipulation of the physical environment to bring about some change of state." Gesture movements can be observed as a hierarchy, as presented by McNeill et al. (1990) (see Figure 2.1).

Arm use and body posture refer to adoption of particular body postures and arm usage patterns. As mentioned by McNeill et al. (1990), Kendon observed stretches where either the left or right arm, or both arms were used, with switches in usage during the interaction. A consistent stretch can roughly be interpreted as a "paragraph", and a shift in usage defines a kinesic unit on this level.

Head movements can appear during a single stretch of consistent arm use and body posture.

Gesture unit is defined as the period between rests of the limbs, starting when the limbs starts moving, and ending when the limbs return to the resting position.

A *gesture phrase* occurs within a gesture unit, and consists of preparation, stroke, and retraction.

Preparation refers to the movement of a limb from its rest position to a position where the gesture begins. *Pre-stroke hold* may occur for a brief period before the actual stroke, after the limb has reached its starting position in the gesture space. McNeill et al. note that the preparation is optional, however, it is rarely omitted. The hold marks the temporary halt in the movement, but without returning to the resting position, which would mark an end of a current gesture phrase.

Stroke is the central part of a gesture phrase and represents the meaning of the gesture, that is the information that is being transferred to the listener. It is synchronized with linguistic segments that may co-appear with the stroke. Additionally, the stroke is usually performed within the central

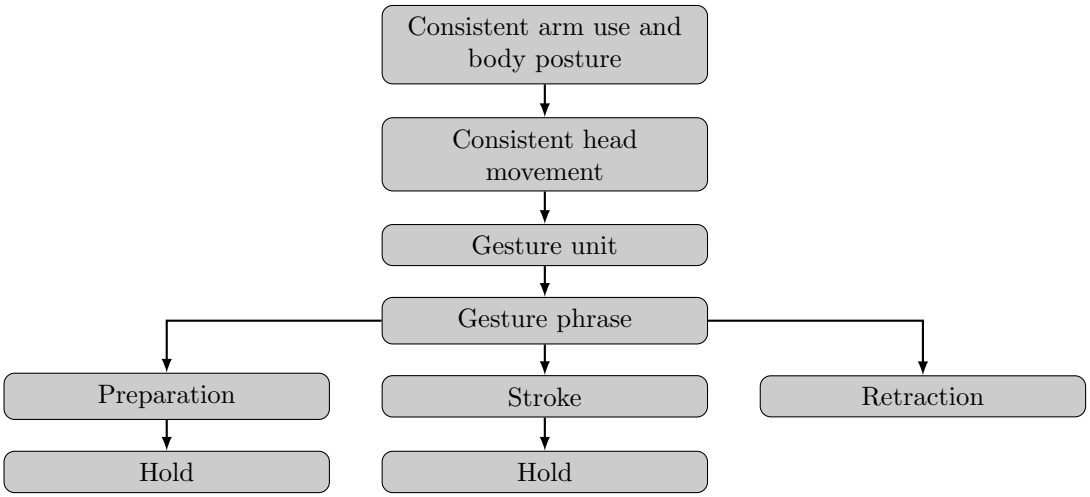


Figure 2.1: Gesture levels, adapted from McNeill et al. (1990)

gesture space, an area bounded with waist, shoulders and arms. A *post-stroke hold* can occur after the stroke, but before the retraction.

Retraction is an optional part of the gesture phrase that is performed to return active hands and arms to their resting positions. In case the current gesture continues into another gesture, this part may be omitted.

Preparation-hold and *Stroke-hold* parts represent hold of active limbs in the current pose for a brief period of time. These two segments usually appear to compensate for desynchronization between co-expressive spoken utterances and gestures.

The focus of the work presented here is recognition of a gesture stroke. The main assumption made is that little information is present in the preparation and retraction, and that therefore the stroke contains the most information of a gesture. This assumption is well supported by McNeill et al. (1990): "While a gesture phrase cannot exist without a stroke, by definition, the other phases are optional. [...] In this book most of our observations refer to the lowest level of the kinesic hierarchy, and within this level, to the stroke phase. [...] The stroke is the phase that carries the gesture content." It directly influences the design of the recognition algorithm and of the training samples. The recognition should be invariant to the actual starting position of the gesture (e.g., a resting position, or a continuation from a previous gesture). Therefore, training samples should be prepared in such way to minimize included motion related to the preparation and retraction.

Kendon (1988) suggests gesticulation "appears as one of the resources a speaker has for conveying meanings. [...] one finds that gesticulation appears to serve a very wide range of communicative functions. It may be used to disambiguate potentially ambiguous words; it may be used as a device to convey aspects of meaning the speaker's words convey only in part, or perhaps not at all; it may be used to complete a spoken utterance, substituting for a segment of the sentence that which is not spoken. In many cases it is possible to see how the speaker deploys gesture together with speech in a way that suggests that he divides the conveyance of his meaning between these two modalities in a way that quite aptly achieves an economy of expression or a particular effect on the recipient."

Furthermore, Kendon (1988) argues that there is "no sharp dividing line [that] can be drawn between spontaneous gesturing that encodes meaning in a holistic fashion and gestures which, like so-called 'emblems,' are not shaped on the spur of the moment but follow an established form within a communication community, or which, like the signs in a sign language, can be shown to

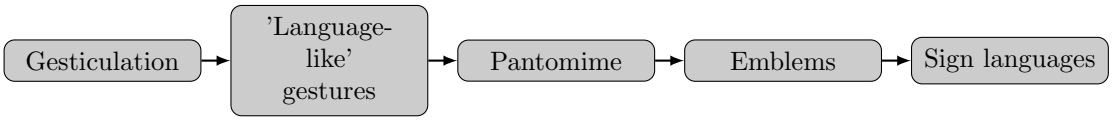


Figure 2.2: Kendon's continuum, adapted from McNeill et al. (1990)

be structured systematically out of recombineable elements and which do indeed refer to meaning units of great generality, as do words." McNeill et al. (1990) named this the "Kendon's continuum" (see Figure 2.2). Moving along the continuum from gesticulation toward sign languages, the co-expressive speech declines, and the language-like properties of gestures and their formalization increase.

Since sign languages posses certain language-like properties, in particular the standardization of sign language signs, their meaning can be more easily obtained. On the other hand, *gesticulation*, *'language-like' gestures* and *emblems* lack standardization, present in sign languages. Hence, the need to define gesture vocabularies, as it was presented in Chapter 4.

With regards to the recognition task, gestures can be static or dynamic. Static gestures represent a specific body posture that is held for a certain amount of time and the meaning of the gesture is held in the particular configuration of relevant body parts. As stated by (Stern et al., 2008), a static gesture is indicated with a fixed position, orientation and configuration, that is having a static posture, while a dynamic gesture is characterized with variations in position, orientation or configuration over time. An "OK" gesture, with a thumb and an index finger forming a ring, while the other fingers are kept extended, is an example of static gestures. On the other hand, the meaning of dynamic gestures is encoded in the motion of some body parts. Greeting someone or attracting someone's attention with a waving gesture is an example of dynamic gestures. Finally, some gestures contain both static and dynamic component, e.g., signs in sign languages.

Task of gesture recognition is to detect these intentional postures or motion and to classify them into particular gesture classes. It has numerous applications, including sign-language for hearing-impaired people, computer interfaces, natural and intuitive human-robot interaction, gaming industry, system remote control, among others. Various tools have been used for gesture recognition, based on the approaches ranging from statistical modeling, computer vision and pattern recognition, image processing, connectionist systems, etc. (Mitra and Acharya, 2007).

Depth image-based sensors have opened up new possibilities in object detection and recognition. Pixels in a depth image carry information about the object's distance from the sensor; hence, the image segmentation is performed taking into account distance gradients. One of the main advantages of this technology is that it is in large part insensitive to illumination changes, which is a major bottleneck for 2D image segmentation methods (Liu and Fujimura, 2004).

Gesture space is the area in front of the gesturing person, in which most of the gestures occur (McNeill, 1992). It can be seen as a shallow disk, with bottom half flattened when the person is gesturing seated. Figure 2.3 illustrates different zones of the gesture space, while Figure 2.4 displays occurrence of different types of gestures in the gesture space, based on gesturing of six persons. Gestures prevalently appear in the frontal plane of the body and occurrences of gestures behind the body plane is rare. Cultural influences can affect the density of occurrence, e.g., Turkana speakers tend to gesture more in the area around the head, compared to the European speakers (as observed on English, French, German, Italian and Georgian speakers). In the cases presented in Figure 2.4, iconic gestures tended to occur more in the central area (Figure 2.4a, metaphorics in the lower central area (Figure 2.4b), deictics in the peripheral area (Figure 2.4c) and beats were grouped in clusters, where each cluster belonged to one person and indicated their "favorite" area for these gestures (Figure 2.4d).

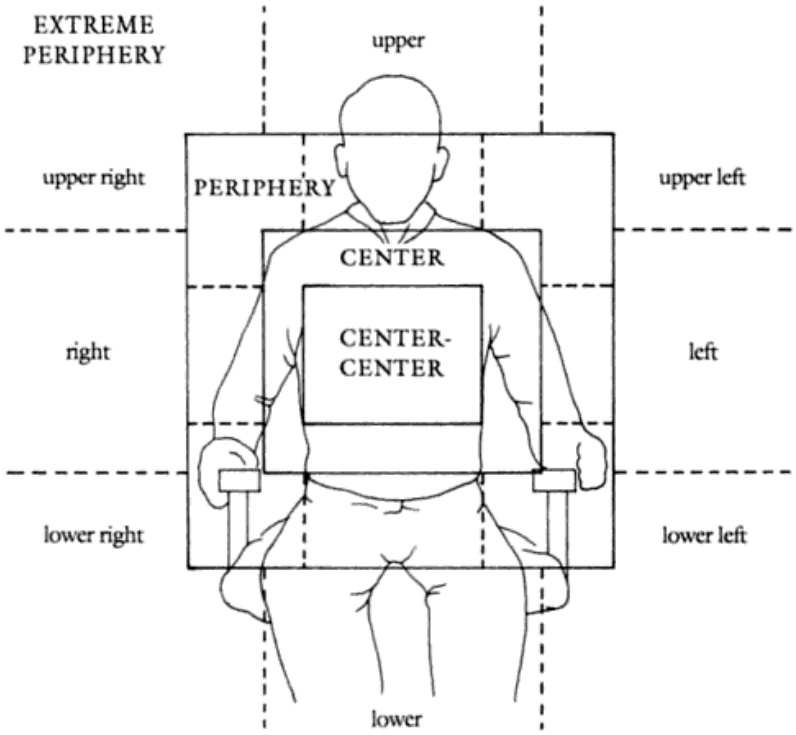


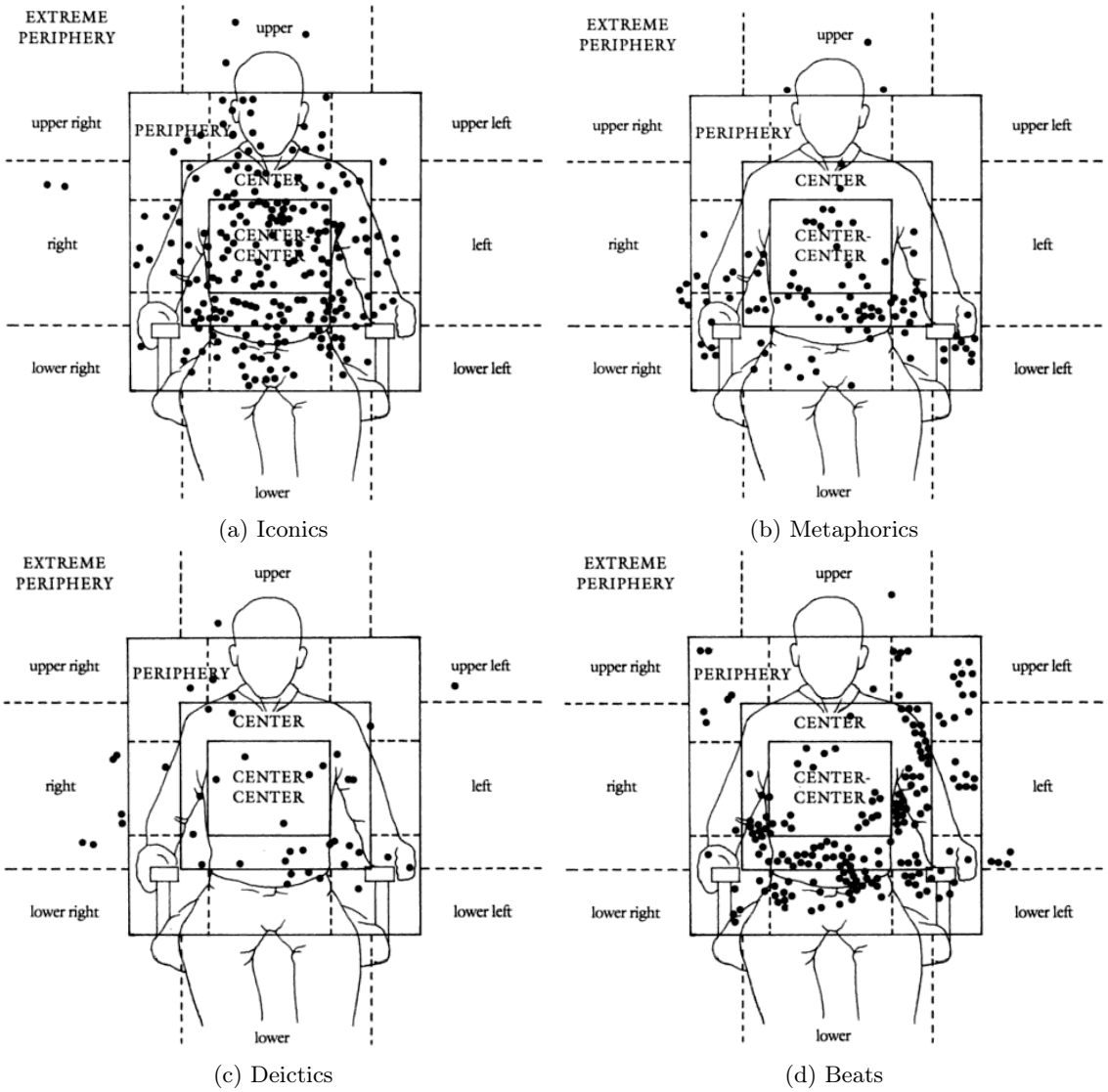
Figure 2.3: Gesture spaces (McNeill, 1992, image taken from Chapter 3)

2.1 Gesture Taxonomies in Human-Robot Interaction and Human-Computer Interaction

This section presents gesture taxonomies described by Quek (1995), Pavlovic et al. (1997), and Nehaniv et al. (2005).

Quek (1995) differentiated gestures into *manipulative* and *communicative* gestures, with regard to their use in human-computer interaction. The purpose of manipulative gestures is object manipulation while communicative gestures are used to transfer information. One qualitative difference is the possibility of visual interpretation. Communicative gestures are visible, providing enough information for their recognition. This means that occluded body parts carry no information required for interpretation. On the other hand, manipulative gestures, e.g. rotation movements of arms while opening a valve, are usually not used for information transfer and, therefore, they might not be fully visible and even interpretable.

Communicative gestures can be *symbols* and *acts*. Symbols are arbitrary in nature. They can be *referential* or *modalizing*. Referential symbols refer to an object or a concept (e.g. holding the index and the middle fingers up as a peace sign). Modalizing symbols influence how a statement is understood (e.g. saying "Have you seen its size?" with the hands wide apart to indicate that the object was large). Acts can be *mimetic* and *deictic*. Mimetic gestures imitate the referred object or its use, such as holding both clenched fists together in front of the body and then pulling them closer, as if they are pulling something with a rope. Deictic gestures are gestures that point to the referent. They can be further classified as *specific* (pointing to a particular object), *generic* (pointing to an object which symbolizes the whole class) and *metonymic* (pointing to an object to signify some other entity related to it).



(a) Iconics (b) Metaphorics

(c) Deictics (d) Beats

Figure 2.4: Frequency of gestures according to their type (McNeill, 1992, image taken from Chapter 3).

Pavlovic et al. (1997) extends the taxonomy by Quek to include unintentional hand movements, such as rubbing hair with the hand, that do not carry any useful information.

Nehaniv et al. (2005) present a gesture taxonomy for application of gestures in human-robot interaction, including *irrelevant* or *manipulative* gestures, *side effect of expressive behavior*, *symbolic gestures*, *interactional gestures* and *referential gestures*. Irrelevant and manipulative gestures correspond to unintentional hand movements and manipulative gestures, as used by Pavlovic et al. (1997). Side effects of expressive behavior represent gestures that can be observed during speech, with no particular information transfer role. Symbols are defined in a similar way as they are defined by Quek (1995). Interactional gestures are used to mediate the interaction between partners. Referential gestures are similar to deictic gestures, as defined by Quek (1995).

The goal of *manipulative* gestures is usually not information transfer, therefore the focus of this work is on *communicative* gestures, according to the taxonomy presented by Quek (1995). Considering classification by Nehaniv et al. (2005), the focus of the work is on symbolic and interactional gestures. Furthermore, the focus is on dynamic uni- or bi-handed gestures, and partially on static gestures, by including dynamic information contained in the preparatory and return movements. While the hand pose provides relevant information for certain gestures (e.g., the mentioned "peace" sign, or a dynamic gesture together with a particular hand posture crucial for correct recognition), the framework does not take it into account.

Deictic gestures represent a special case. The primary goal of deictic gestures is to manipulate the attention of an interacting partner to an object or an event of interest, as reported by Masataka (2003). The proposed framework does not handle pointing gestures. However, they have a characteristic signature, containing of the preparatory movement, the hold, during which the finger is still and pointing to a location of interest, and the retraction, taking the hand to the resting position, or preparing for the next gesture. The framework could be extended to include additional branch for processing of pointing gestures, in case they are detected based on their signature.

Finally, the choice of these classes is supported by the analysis presented by McNeill (1992). He presented a frequency of gesture types during narration. In summary, a third of narrational clauses were accompanied with iconic gestures, while another third were accompanied with beats. A quarter of the clauses did not have co-occurring gestures, and a smaller part were occurrences of metaphorics and deictics.

Chapter 3

Attention Systems for Human-Robot Interaction

3.1 Overview

Evidence from developmental psychology studies show that the development of skills to understand, manipulate and coordinate attentional behavior is a requirement for imitation learning and social cognition. Pointing is a way for manipulating the attention of someone else. It is not yet clear whether this behavior is innate, learned by imitation, or if it results from reaching behaviors, in its first developmental stage. Recognizing and performing pointing gestures is very important for being able to share attention with another person (Kaplan and Hafner, 2006). In this chapter, an implementation of an attentional model for a humanoid robot is presented, employing both a mechanism for saliency detection and a mechanism for pointing gesture generation (Hafner and Schillaci, 2011).

Moreover, the ability to distinguish between animate and inanimate objects (implemented here by a motion detection filter) and whether this object is a person to interact with, as well as the ability to detect and follow eye gaze (which requires the recognition of faces) are essential prerequisites in the development of a Theory of Mind – or Mindreading (Baron-Cohen, 2001) – the ability to understand the motivations and desires of other people which identifies the clear distinction between us, humans, and other animals.

The robot is also equipped with something similar to human focused attention. It is not hard to imagine an everyday situation during which a person focuses their attention on a certain aspect of the environment, e.g. talking to someone, reading an interesting article or doing work. During the time when their attention is focused on these processes, they suppress most other salient events. This is modeled with a change from exploration to interaction phase once certain conditions have been satisfied, and it is explained later in more detail.

Section 3.3 of the chapter presents the notion of an attentional model and prior work, and a developed implementation of an attentional model in a humanoid robot, using an egosphere as a model of memory representation. This attentional model enables the robot to explore its near environment and locate and remember important detected features, such as motion and faces. Four different behaviors were implemented in order to realize an interaction game. It was tested through an interaction game with participants and the Godspeed questionnaire was used to measure how the participants perceived different behaviors of the robot.

The ability to share the attention with another individual is essential for having intuitive interaction. Two relatively simple, but important prerequisites for this, saliency detection and

attention manipulation by the robot, are identified. By creating a saliency based attentional model combined with a robot ego-sphere and by adopting attention manipulation skills, the robot can engage in an interaction with a human and start an interaction game including objects as a first step towards a joint attention.

An interaction experiment is proposed in which participants can physically interact with a humanoid robot equipped with mechanisms for saliency detection and attention manipulation. The experiment tests four combinations of activated parts of the attention system, resulting in four different behaviors.

The aim is to identify those physical and behavioral characteristics that need to be emphasized when implementing attentive mechanisms in robots, and to measure the user experience when interacting with a robot equipped with attentive mechanisms.

Two techniques are adopted for evaluation of saliency detection and attention manipulation mechanisms in human-robot interaction – user experience, measured by qualitative and quantitative questions in questionnaires, and proxemics, estimated from recorded videos of the interactions.

Section 3.4 discusses the main outcomes of the work and states the main conclusions presented in this chapter.

This chapter is partially based on work presented in Bodiroza et al. (2011), Schillaci et al. (2013).

3.2 Background

Gestures are important in human-robot interaction, as they can enable intuitive and natural interaction. However, interaction relies on the ability to share attention. Kaplan and Hafner (2006) identified the prerequisites for joint attention. Being able to detect, locate and attend to a prominent feature in environment is an important prerequisite, since every interaction relies on presence of an agent with whom another interacts. As Kaplan and Hafner (2006) pointed out, there are multiple open questions to development of joint attention. This chapter addresses the issue of equipping the robot with an attentional model that will enable it to explore its immediate environment and find interesting features.

Social robotics, a new area in human-robot interaction has seen an increase in research in the past decades. As a result, there is a shift in a typical robot working environment from industrial halls to everyday places, such as homes (Jones, 2006), workplaces and service business, e.g., a museum tour-guide robot (Thrun et al., 2000) and a guide robot in a shopping mall (Kanda et al., 2009). As a result, a requirement for natural and intuitive human-robot interaction has emerged.

Current social robotic systems usually rely on interaction protocols which decrease the intuitiveness of the interaction itself, causing frustration and despair in the user. Recently, interest has been focused on measuring the efficacy of robot behaviors and its perceived intelligence based on the evaluation from human users (Burghart and Steinfeld, 2008). Indeed, measuring human-robot interaction could suggest what and how to improve in the cognitive abilities and in the appearance of the robot.

When human-robot interaction fails, the reason most often lies in the fact that the robot and the human try to communicate about different things and that the human partner has wrong expectations of the robotic partner. Several prerequisites have been identified (Kaplan and Hafner, 2006; Schillaci and Hafner, 2011a) about both physical and cognitive features that let a robot interact effectively and naturally with a human user.

Two skills that can be given to the robot and in combination already result in interesting interaction behavior are the ability to focus the attention on salient features and the ability to manipulate each other's attention by looking and pointing.

There are different existing approaches for saliency detection and visual attention. One approach relies on using saliency maps, also known as attention activation maps, proposed by Itti and Koch (2001), based on the Feature-Integration theory (Treisman and Gelade, 1980). In this method, an image is analyzed using pre-attentive bottom-up processes. Later on, this model has been extended to include motivation-influenced top-down processes, like in the work of Frintrop et al. (2005). First, the image is analyzed in parallel for particular low-level features, such as color, orientation, motion and so on. However, these features are also influenced by a top-down process, which is motivation-specific, as well as depending on the current context, e.g. searching for faces or specified objects. Therefore, regions detected by the bottom-up process are analyzed for specific features. The result of these processes is a saliency map, where salient regions are represented with higher peaks in the map, and their height (or intensity) represents their saliency values. This is in accordance with Treisman (1985), who states that the visual perception process is functionally divided in two levels: early pre-attentive and later attentive. During the pre-attentive stage a rapid scanning of the image is performed, while in the attentive stage the attention shifts toward selected regions.

Another approach partly relies on saliency maps, but also introduces the concept of a multimodal salient ego-sphere (Peters et al., 2001; Fleming et al., 2006). It is a tessellated sphere, where each of the nodes (vertex on the tessellated sphere surface) represents one point in space. Input images are analyzed by using the aforementioned saliency detection method or by other methods such as face detection (Viola and Jones, 2004) for generation of saliency maps. The mean value of salient regions from the resulting maps are then assigned to the nodes, which are closest to the real position of the projection of the salient region to the sphere surface.

Ruesch et al. (2008) described a framework which is based on the previous concept. However, due to the higher computational capacity of the iCub robot, they use a matrix projection of the ego-sphere, which leads to higher precision of the perception of the world. This also increases the computational complexity, due to required image transformations and a higher number of arithmetic operations required per iteration.

Here the emphasis is placed on the fact that robots need to reach joint attention with the users for having successful interactions. This has not been achieved so far, since joint attention not only requires attention on the same features in the environment, but also skills in attention detection, attention manipulation, social interaction skills and even intentional understanding (Kaplan and Hafner, 2006). In other words, interacting partners need to share their attention on a particular feature, coupled with intention detection and understanding. Features can be physical (e.g. visual or auditory), as well as abstract (e.g. mental imagery). Without joint attention a robot will not be able to achieve a degree of interaction that could be compared to a interpersonal interaction. That is, the robot would still need to rely on receiving precise commands from the interacting participant, which would relate to the actions that it could automatically perform.

Several metrics for measuring HRI have been proposed, from measuring the ability of a robot to engage in temporally structured behavioral interactions with humans (Jonsson and Thorisson, 2010), to evaluating robot social effectiveness from different points of view (engineering, psychological, sociological) (Steinfeld et al., 2006). A series of metrics was adopted for the survey, based on cognitive science studies about measuring social skills in humans and based on studies about how robots are perceived by humans and whether this perception affects the expectation humans have about robot intelligence, the so-called Godspeed questionnaire (Bartneck et al., 2009).

Quantifying human behavior usually requires joint analysis of video recordings, questionnaires and interviews. In this work, the first two methods are used for quantifying the quality of robot behavior. Four interaction experiments between a humanoid robot and a user were setup and recorded. After each experiment, the user was asked to fill a questionnaire on the quality of the interaction and on the perception of several functional and physical properties of the robot. To the best of author's knowledge, very few studies have been done so far on correlating human perception

of robot skills (measured with the Godspeed questionnaire, whose reliability was tested) with proxemic distances. For example, Takayama and Pantofaru (2009) adopted part of the Godspeed questionnaire in their measurements, finding that people who held more negative attitudes toward robots felt less safe when interacting with them. They also studied human personal space around robots, finding that experience with owning pets decreases the personal space that people maintain around robots, experience with robots decreases the personal space that people maintain around robots, and a robot looking at people in the face influences proxemic behaviors. The latter suggests to perform proxemics analysis when measuring attentive mechanisms in robots.

By choosing a system which resembles the human visual attention system, it can easily be that the expectations of the interacting person will have a higher satisfaction rate. In this case the user might perceive the interaction to be more intuitive and natural.

3.3 Implementation of an Attentional Mechanism for a Humanoid Robot

In this section, an implementation of an attentional model for a humanoid robot is described, encompassing two fundamental skills for joint attention. Furthermore, the quality of this implementation is evaluated through a user study. By evaluating robot skills, the goal is to identify those characteristics that need to be emphasized when implementing attentive mechanisms in robots and to identify correlations between them.

3.3.1 Saliency Detection and Attention Manipulation

Attention is a cognitive skill, studied in humans and observed in some animal species, which lets a subject concentrate on a particular aspect of the environment without the interference of the surrounding. There is evidence from developmental psychology studies that the development of skills to understand, manipulate and coordinate attentional behavior lays the foundation of imitation learning and social cognition (Tomasello, 1995).

In our world, we are constantly surrounded with items, such as objects, people and events, which stand out to their neighboring items. This is represented with the saliency of those items. Saliency detection represents an attentional mechanism, through which those items are discovered, and it enables humans to shift their limited attentional resources to those objects that stand out the most.

There are two approaches that can be combined – a bottom-up, pre-attentive process and a top-down process influenced by motivation. Bottom-up detection uses different low-level features (e.g. motion, color, orientation and intensity) for saliency detection. Top-down detection relies on high-level features, and it is highly influenced by our current goals and intentions. In this case, the motivation system was partially preprogrammed to show how different behaviors can result in the activation or deactivation of parts of the attention system, actually implementing a top-down approach for saliency detection, or in the activation of attention manipulation. The combination of bottom-up and top-down processes is highly inspired by similar mechanisms in humans (Itti and Koch, 2001; Treisman, 1985).

Figure 3.1 gives an overview of the attention mechanism implemented on the humanoid robot Nao.

Saliency Detection

Saliency detection represents a process of image analysis during which regions of interest are detected in the scene. The implemented approach employs bottom-up, based on low-level feature

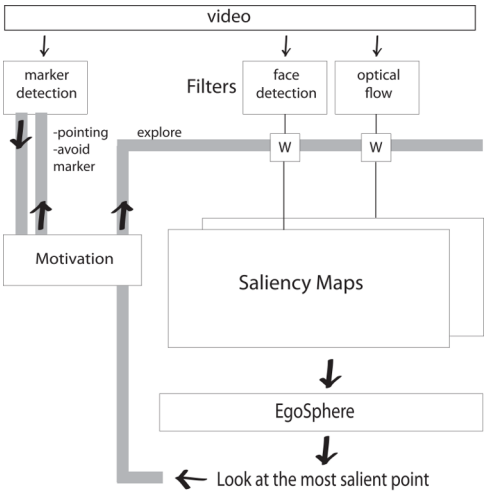


Figure 3.1: Overview of the attentive mechanism. Frames are analyzed by three different filters which are activated by the motivation system. Optical Flow and Face Detector filters feed the ego-sphere, while the marker detector filter stores objects in a different vector. The motivation system activates or deactivated filters and movements according to its current state.

detection, and top-down processes, based on current motivation, partially reproducing similar mechanisms in humans (Itti and Koch, 2001; Treisman, 1985).

Three filters are used in the current setup: *motion detection*, *face detection* and *object detection*. Motion detection is performed using Farnebäck's optical flow filter. Face detection is performed using the method developed by Viola and Jones (2004). AR markers are used as a shortcut for simplified object detection.[1] Information coming from face and motion detection filters are stored in an ego-sphere, keeping track from which of those two channels information come from. The robot directs its attention to the point which has the highest saliency. In the current implementation, AR marker data are stored separately outside the ego-sphere.

Ego-Sphere. Saliency ego-sphere is a multi-modal, egocentric, spherical map, which represents salient areas in the robot's surrounding. It enables a robot to shift its attention from one salient area to another one in an apparently random and natural fashion.

The sphere is centered at the robot's neck coordinate system, while the saliency map of its surrounding is projected on the sphere's surface. Through mechanisms of habituation, inhibition and forgetting of salient areas the robot is able to explore its surroundings, and by finding areas of maximum saliency, it locates the next area to be attended. Furthermore, the robot uses pointing to point to the currently attended object, as a first step towards joint attention (Kaplan and Hafner, 2006).

There are different representations of the ego-sphere. One representation is where full salient maps of different scenes are stored in a matrix, which represent spherical surface and where images are aligned to a certain extent (Ruesch et al., 2008). Another approach is to reduce the projection and search space by tessellating the sphere and storing information about salient areas in the edges of the tessellated sphere (Fleming et al., 2006), as illustrated in Figure 3.2. The latter approach introduces errors in projection of salient areas, because projection space is reduced to a set of nodes on the sphere surface. Additionally, the mean saliency of a salient area is computed and assigned

[1] http://www.hitl.washington.edu/artoolkit/

to the closest point, which is found using nearest neighbor search. While the former approach has higher precision and finer representation of salient areas, the latter approach is faster, due to the lower number of arithmetic operations that need to be performed during the projection and the search. Due to Nao's computational limitations, the ego-sphere is represented with a tessellated sphere, where information about salient areas is stored in the edges of the sphere.

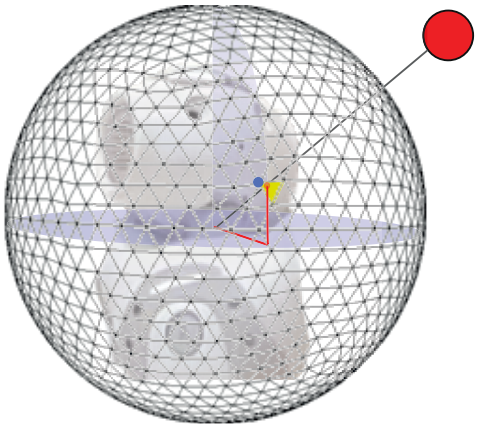


Figure 3.2: Illustration of an ego-sphere representation.

The sphere tessellation is performed by recursive division of triangle faces of an icosahedron. By increasing the recursion depth, which represents the number of recursive calls to the tessellation function, each initial face is divided into a higher number of smaller faces, achieving a higher number of nodes and smaller error of projection of salient nodes. The relation between the recursion depth and the tessellation frequency, used by Peters et al. (2001), is $f_t = {d_r}^2$, where f_t is the tessellation frequency and d_r is the recursion depth. In essence, the tessellation frequency represents the number of edges which connect centers of two neighboring pentagons. The implementation uses a recursion depth of 4, resulting in 2562 nodes, with the mean theoretical projection error being 1.15° and maximum 2.65°.

Habituation, inhibition and forgetting. These processes are employed in order to favor shift of attention to information in new locations. This is inspired by the inhibition of return mechanism in humans (Posner et al., 1985). Habituation is the process during which the robot gets used to the attended point, which results in loss of interest in that point. It is modeled with the following function:

$$h(t) = h(t-1) + w_h(1 - h(t-1)), \tag{3.1}$$

where $w_h \in [0, 1]$ represents the habituation weight.

When habituation to a certain salient point exceeds a predefined habituation threshold level, inhibition is turned on with a change from 1 to 0, which results in the appearance of a shift of attention to the next most salient point. The reported saliency of a point is the product of its current inhibition value and its saliency.

However, inhibition has a temporary nature and it is updated with an inhibition weight, according to the following function:

$$i(t) = i(t-1) + w_i(1 - i(t-1)), \tag{3.2}$$

where $w_i \in [0, 1]$ represents the inhibition weight.

Habituation and inhibition weights affect the speed of the respective processes. For certain filters, such as motion, locations of salient regions will quickly change in time. Motion can be used to initially draw attention, but it should also decay faster, because it is only present for a short period of time. Other filters can be employed to detect different salient objects in and around these areas, such as face or color detection. Under the assumption that the face does not move much during the interaction, a robot should have slower habituation to locations that contain faces.

The values of salient locations decay according to the decay factor $d \in [0, 1]$, which enables a form of short-term memory for the robot. After each iteration, each salient location will have a lower value than before, thus newer information will have higher saliency than older information.

Attention Manipulation

The ability to draw attention of others is compulsory for having a natural and effective interaction. Pointing gestures can be seen as one of the approaches for this attention manipulation. Therefore, recognizing and performing pointing gestures is very important for being able to share attention with another person, as joint attention relies on the ability of the person to catch the interest of its interlocutor toward an object (Kaplan and Hafner, 2006).

Looking back and forth between the person and the object, or pointing toward the object are examples of non-verbal attention manipulation behaviors. The following subsections present how pointing gesture capabilities in a humanoid robot were implemented.

Pointing. Expression of pointing gestures has also been observed during the first 18 months, but the course for becoming a fundamental communication tool can be divided into two phases, as shown in the following subsection.

Already at the age of 9 months, infants are able to use imperative pointing gestures (Baron-Cohen, 2001). This form of pointing is exhibited by the child regardless of whether an adult is present in the room or is actually looking at the child. Since imperative pointing is not directly used to draw attention, it is possible that it arose from grasping objects within the reach of the child and turned into pointing for objects that were outside the field of grasp.

Some studies showed that there is no relation between producing pointing gestures and understanding them (Desrochers et al., 1995). This hints to the conclusion that the two skills develop independently from each other and strengthens the hypothesis that pointing may arise from grasping and is not learned by imitation. Schillaci and Hafner (2011b) tested this hypothesis, through a robotic experiment where a humanoid robot learned to reach objects (or positions in its action space) by building a body map during random body babbling.

Body Babbling. Body babbling observed in infants has been classified by Meltzoff and Moore (1997) as a mechanism that provides experience for mapping movements to the resulting body configurations. Schillaci and Hafner (2011a) used analogous motor babbling for learning the mapping between different sensory modalities and for equipping the robot with predicting abilities of sensory consequences (in this case, the position of the hand of the robot) from control commands applied to its neck and its arm. Learning through self-exploration was implemented on a humanoid robot, whose dimensions resemble those of a child, actually simulating the real visual input perceived by a young human subject (Schillaci and Hafner, 2011b).

The robot was equipped with an elementary attentive system for perceiving parts of its own body. During babbling, the robot performed random arm movements and tried to follow them moving its head according to the attentive system. [2]

[2] These behaviors were implemented on the humanoid robot Nao. For learning, four degrees of freedom of the

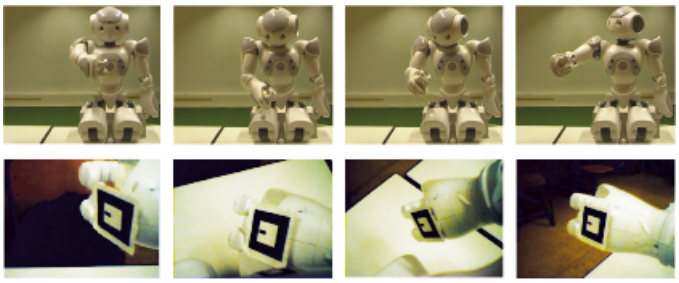


Figure 3.3: A typical babbling sequence using the Nao platform. In the lower part are the corresponding frames grabbed by the on-board camera (note that the camera is placed below the fake eyes of the Nao).

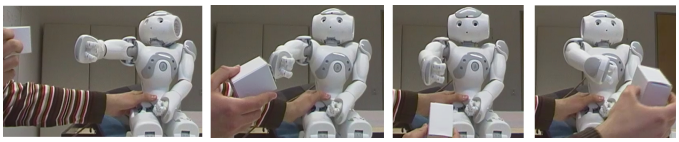


Figure 3.4: Example of a sequence of pointing gestures performed by the robot, generated with the learned model.

Predictive Model. Proprioceptive data (joints angles) and external data (visually estimated position of the hand of the robot) collected during babbling were stored into a knowledge base for feeding a simple predictive model used for anticipating visual consequences of motor commands applied to the arms and to the neck of the robot or, inversely, for solving the inverse kinematics problem (that is, for generating joint configurations of the arm in order to reach a given point in the action space of the robot).

Forward predictions have been performed using a k-Nearest Neighbors based algorithm, as explained by Schillaci and Hafner (2011a)).

The pointing experiment described by Hafner and Schillaci (2011) was implemented using the predictive model for generating joints configurations of the arm of the robot that result in placing its hand as close as possible to an object tagged with a fiducial marker and situated outside the field of reach of the robot. This resulted in pointing towards an object, when it was outside the reach of the robot (Hafner and Schillaci, 2011). Thus, it was demonstrated that the system used for reaching points in the action space of the robot's arm can be used for generating pointing gestures outside of its field of reach. Figure 3.4 shows the robot performing a sequence of pointing gestures toward an object tagged with a marker.

Parameters

The following parameters for the algorithm were chosen so that overall the ratio of accuracy and speed is optimized.

Camera resolution and image scale factors. The camera resolution is set to 320×240 pixels. However, the image is resized for different filters. In the case of face detection in a typical experimental setup, there is enough level of detail after scaling down by 2. The image for motion

arm, two degrees of freedom of the neck, together with the estimated position of the hand (extracted from the video grabbed from the bottom camera of the robot) were used. The hand was tagged with a fiducial marker, which allowed an easy and fast estimation of its position both in the image and in the frame of reference.

detection is scaled down by 4, which is inspired by the human peripheral vision. Human peripheral vision is blurred, but it provides good results for motion detection.

Minimum blob size for motion detection. Some parameters were adopted and tuned in order to avoid tracking motion caused by noise. Dense Optical Flow has been computed using the Farnebäck algorithm, which outputs an image with the same size as the original image containing the magnitude of the movement for each pixel of the image.

First, a minimum flow magnitude threshold is set in order to prevent adding points to the ego-sphere generated by noise. The value was empirically set to 7. (velocity in pixel). Then, blobs were detected in the flow magnitude image, taking into account only blobs bigger than a given size (50 pixel, in a 80×60 image), in order to cut out isolated moving pixels or small areas probably generated by noise.

Minimum face size and scale factor. Minimum face size for the Viola-Jones algorithm is set to 35×35 pixels and the scale factor is set to 1.4 (faces detected up to $1.5m$).

Habituation and inhibition weights, habituation threshold level, decay step for forgetting. These parameters influence for how long the robot will focus on one salient area and how fast it will forget older salient areas. They were chosen so that the robot seems responsive enough in a typical interaction. The habituation weight, w_h, is lower for the face detection compared to the motion detection. For the inhibition weight, w_i, the relation is reversed. This results in a behavior where the robot habituates faster to motion, since it is present for a short amount of time at one location, while the face usually does not move much in the usual direct interaction. In the current implementation, weights are $w_{h,face} = 0.4$, $w_{h,motion} = 0.7$, $w_{i,face} = 0.5$, $w_{i,motion} = 0.15$, and the decay factor is $d = 0.2$.

3.3.2 Behaviors

A partially preprogrammed motivation system was implemented by which the robot can change its behavior due to its current beliefs and desires. Four different behaviors were implemented, as explained in the next subsections. This motivation system was implemented to show how each behavior can result also in the activation or deactivation of part of the attention system, actually implementing a top-down approach for saliency detection.

Exploration

In this motivation state, all the saliency filters are activated, the ego-sphere is updated frame by frame, and control commands are applied to the joints of the head to let the robot focus on salient events. As depicted before, the robot is attracted by movements, faces and objects.

Interaction

In the interaction phase, the robot is not exploring anymore using face and motion filters. Its behavior is now focused on looking and pointing at the object (such movements are generated using the predictive model explained before).

Interaction avoidance

In this state, the robot just detects markers and, if any, moves its head toward a configuration far from the current one, actually trying to look towards areas which do not contain any markers.

Figure 3.5: Nao with colored stickers that indicate the position of the fake eyes.

This behavior has been implemented for trying to convince the user that the robot is bored and it does not want anymore to follow the interaction session.

Full Interaction

This behavior is composed of a sequence of the previous behaviors. The first performed action is *exploration*. Once the robot has detected a person to interact with and an object which can be used to draw the attention of the user, its motivation state changes to *interaction*. The interest value decreases over time and it specifies the lapse of time the robot stays in the interaction state. It was observed in a previous interaction experiment that users used to bring the object and hand it to the robot for the whole session. Using this interest decay variable, the interaction phase can be interrupted and the robot's behavior can be modified to *interaction avoidance*. The interest variable (initialized as 1) decrease slowly (by 0.005) when the person is handing the object to the robot or rapidly (by 0.025) if not. Motion of the marker is used as the indicator that a person is holding the object with the marker.

When the interest factor goes below zero, the robot would shift its behavior to *interaction avoidance*. After a while, the current behavior is set back to *exploration*.

3.3.3 Robot Platform

The robot platform is the Nao from Aldebaran, a humanoid robot around $57cm$ tall. For the experiment, only the degrees of freedom in the arms and the neck were used. The lower camera is positioned below two eyes, which resulted in the robot not seeing an object if it is brought close to the eyes. For that reason two fake eyes were placed on the sides of the lower camera, and the real eyes were covered with a tape, as shown in Figure 3.5. Figure 3.6 shows a typical setup during the experiment.

The attention mechanism was implemented in C++ using the framework of the Nao Team Humboldt (Burkhard et al., 2010). The attentional mechanism is fully executed on-board the robot and there is no remote processing of the data. The robot is connected to the computer through Ethernet. A robot control program is running on the computer which is used to visualize the data and activate required modules for the attentional mechanism in the framework.

Such robot was adopted for measuring the users' expectations about the robot's skills due to its anthropomorphic form. Moreover, its small child-resembling size could reduce users' expectations, thus increasing the positive evaluation of the interactions.

The robot has limited computational resources. The implementation described here, at the current state, lets the robot process all the filters at a rate of approximately 7-8 frames per second. The computationally most expensive algorithms are those related to image processing, e.g. the face detection filter and the optical flow filter, which together take almost 110 ms per calculation. This results in slower movements and reactions when the robot is in the exploration state and in the exploration part of the full interaction state, which could affect the intuitiveness of the interaction. However, in a preliminary experiment, the participants rated the speed of robot as good. An interesting research question could ask what is the proper movements speed a robot might exhibit in order to be perceived as not dangerous or with good reaction times. This topic is included in the future development of this work.

Furthermore, in the current experiment, although the fastest processing was in interaction avoidance, people perceived the robot as less responsive than during interaction and full interaction.

3.3.4 Experimental Setup

For testing the robot's abilities on saliency detection and attention manipulation, set up an interaction experiment is set up in which participants could physically interact with the Nao robot equipped with the mechanisms previously shown. After the interaction, participants were asked to fill in a questionnaire about their perception of the interaction (see subsection 3.3.4).

The proposed experiment aimed at several goals: test the quality of the implemented saliency detection and attention manipulation mechanisms; identify those physical and behavioral characteristics that need to be emphasized when implementing attentive mechanisms in robots; measure the user experience when interacting with a robot equipped with attentive mechanisms; find correlations between heterogeneous robot features perceived by the participants during the exhibition of attentive mechanisms; analyze the differences in the perception depending on the different behaviors performed by the robot.

An overview of the attentive mechanism was presented before in Figure 3.1. Frames are analyzed by three different filters which are activated by the motivation system. When the current state of the robot is *exploration*, all the filters are active. Optical Flow and Face Detector filters feed the ego-sphere, while the marker detector filter stores objects in a different vector. When both a face and an object have been detected, the behavior of the robot changes to *interaction*, deactivating the ego-sphere and activating the pointing gestures. When the interest on interaction is low enough, the behavior changes to *interaction avoidance*. The marker detector is still active, because the robot just generates head movements in order to avoid the sight of the object (for letting the user understand that the robot is not interested in the pointing interaction anymore).

The implementation was tested on the four behaviors, as described previously in Subsection 3.3.2.

Hypotheses

Several outcomes of the experiment were expected. It was expected that the level of interactiveness of the robot was positively correlated with the level of excitement and perceived intelligence.

Figure 3.6: Experimental setup showing interaction between the Nao and a person.

Playing with the robot in the interaction state might be more exciting and satisfactory than playing with it in the avoid interaction one.

Multi-modal interaction (through arms and head movements) might increase the perception of interactiveness. Similarly, a less interactive behavior might decrease user satisfaction and cause the participants to behave nervously.

Anthropomorphic attributes might be positively correlated with the perception of intelligence.

Hafner and Schillaci (2011) demonstrated how pointing gestures can emerge from reaching behaviors. In a preliminary experiment, participants were asked how they interpreted the movements of the robot performing the interaction behavior, expecting that reaching commands can be perceived as a desire to grasp the object.

Procedure

The experiments consisted of the robot performing the behaviors described in the previous section in four separate interaction sessions, one per each of the four behaviors. The experiment supervisor manually activated or deactivated them. Figure 3.6 shows a frame taken from a typical interaction session. The user sat in front of the robot at a distance of ca. 90 *cm*. For each person, each interaction test lasted one minute. The interaction was recorded with a standard camera (resolution 640×480) placed at ca. 2 meters perpendicularly to the robot-user axis. Beside the table where the robot was standing there was a scale drawn on a whiteboard for the visual estimation (estimated average error: $5cm$) of the distance between the nose of the user and the head of the robot and from the hand of the user and the head of the robot; according to the type of interaction, it was noticed that the users move their hands closer to the robot.

After each of the four interaction sessions, the participants were asked to fill a questionnaire about the quality of the interaction with the robot and about the perception of robot behaviors.

Participants

In total 28 people participated in the survey, which results in a total of 112 questionnaires (four questionnaires per participant, one for each interaction). Some participants missed to answer some questions, but those were only a few questions. It is interesting to note that few participants had negative or neutral responses in all four experiments, regardless of the experiment, together with comments saying that Nao did not want anything because it is a machine. This might be perceived as a negative bias towards robots.

Of 28 participants, 8 were female (28.57 %) and 20 were male (71.43 %). There were 17 Germans, 2 Italians, 2 Serbians, 2 Poles, 1 Czech, 1 Dutch, 1 Estonian and 1 French. Regarding previous experience with robots, 25 persons (89.29 %) had none and 3 (10.71 %) had previous experience –

one with industrial robots, one with Aldebaran Nao and one with Lego Mindstorms. The average age of the participants was 28.12 ($\sigma = 5.64$). Among the participants, 75 % had university level education and 25 % had high-school level education.

Unfortunately, not all the participants allowed us to film their interaction because of privacy reasons (even though they were informed that the data will be kept anonymous and videos will not be published against their will). The video database is composed of 10 videos for *exploration*, 7 for *interaction*, 8 for *avoid interaction* and 9 for *full interaction*.

Measurements

Only recently, performance criteria different from those typical for industrial robots have been adopted for measuring the success of social and service robots. Current criteria lie within the satisfaction of the user (Bartneck et al., 2008).

Two techniques were adopted for evaluating the interaction: questionnaires and proxemics estimated from recorded video sequences of the interaction. So far, the wish was to adopt only metrics related to socio-cognitive skill perception instead of measuring the affective state of the user through the use of physiological sensors.

Questionnaires. A qualitative, anonymous survey was conducted to evaluate how people perceive their interaction with the Nao. Questionnaires are often used to measure the user's attitude. The first problem was related to what type of questionnaire to adopt. Developing a valid questionnaire can take a considerable amount of time and the absence of standardization makes it difficult to compare the results with other studies. That is why a standardized measurement tools for human-robot interaction were adopted in the experiment, in addition to some metrics that were found interesting for this research. The Godspeed questionnaire (Bartneck et al., 2009) was adopted as a part of the survey, which uses semantic differential scales for evaluating the attitude towards the robot. Such a questionnaire contains questions (variables) about five concepts (latent variables): Anthropomorphism, Animacy, Likeability, Perceived Intelligence and Perceived Safety (for a detailed description and for the set of questions, please refer to Bartneck et al. (2009)).

Anthropomorphism refers to the attribution of human features and behaviors to non-human agents, such as animals, computers or robots. Anthropomorphism variables were (left value scored as 1, right value scored as 5): fake – natural, machinelike – humanlike, unconscious – conscious, artificial – lifelike, moving rigidly – moving elegantly.

Animacy is the property of alive agents. Robots can perform physical behaviors and reactions to stimuli. The participants' perception about robot animacy can give important insights for improving robot skills. Variables were: dead – alive, stagnant – lively, mechanical – organic, artificial – lifelike (different from the one in anthropomorphism, as related to the animacy), inert – interactive, apathetic – responsive.

Likeability may influence the user's judgments. Some studies indicate that people often make important judgments within seconds of meeting a person and it is assumed that people are able to judge also a robot (Bartneck et al., 2009). Likeability variables were: dislike – like, unfriendly – friendly, unkind – kind, unpleasant – pleasant, awful – nice.

Perceived Intelligence is one of the most important metrics for evaluating the efficacy of the implemented skills. It can depend on robot competence, but the duration of the interaction is also one of the most influencing factors, as users can become bored if the interaction is long and the vocabulary of the robot's behaviors is limited. Variables were: incompetent – competent, ignorant – knowledgeable, irresponsible – responsible, unintelligent – intelligent, foolish – sensible.

Perceived Safety is a metric for estimating the user's level of comfort when interacting with the robot and the perception of the level of danger. Variables were: anxious – relaxed, agitated – calm, quiescent – surprised (this variable was recoded, as explained in the next paragraph).

The reliability of the questionnaire was analyzed by the authors of the Godspeed, who claim that such questions have sufficient internal consistency and reliability; to confirm this, Cronbach's alpha[3] was computed for each latent variable again. It was found that Cronbach's alpha was negative ($\alpha = -1.111$) for the latent variable Perceived Safety, due to a negative average covariance among items. This violated reliability model assumptions for that set of variables, due to a miscoding of a variable. In fact, the questionnaire is written in such a way that high values of one variable mean the same thing as low values of the other variable; the miscoded variable was: Quiescent (scaled as 1) to Surprised (scaled as 5), probably due to the fact that participants intended quiescence as a synonym for calmness (the previous variable was Agitated, coded as 1, or Calm, coded as 5). After recoding the quiescent – surprised variable, the Cronbach's alpha proved to be higher ($\alpha_{PerceivedSafety} = 0.839$)[4]. No other problems were found with the rest of the latent variables: $\alpha_{Anthropomorphism} = 0.825$, $\alpha_{Animacy} = 0.853$, $\alpha_{Likeability} = 0.813$, $\alpha_{PerceivedIntelligence} = 0.750$.

In addition to the Godspeed questionnaire, a new latent variable for measuring the concept of User Satisfaction was introduced, with two variables: frustrating – exciting and unsatisfying interaction – satisfying interaction (high Cronbach's alpha: $\alpha_{UserSatisfaction} = 0.799$).

Open questions were also introduced about the understanding of the behavior of the robot, its desires, its aiming to interact or not, its successfulness, its gender (with the explanation of the chosen one), its age, type of communication during the interaction, expectations about future improvements and differences between Nao and humans.

Proxemics. According to the sociological concept of proxemics, humans, as well as animals, use to define personal spheres which delimit areas of physical distance that correlate reliably with how much people have in common (van Oosterhout and Visser, 2008). The boundaries of such spheres are determined by factors like gender, age and culture. Coming inside the sphere of another person may let him/her feel intimidated, or staying too far can be seen as cold or distant. Four spheres were identified, according to van Oosterhout and Visser (2008): Intimate Distance (from 0 to 45 cm), reserved for embracing, touching, whispering; Personal Distance (from 45 to 120 cm), reserved for friends; Social Distance (from 1.2 to 3.6 m), reserved for acquaintances and strangers; Public Distance (more than 3.6 m), reserved for public speaking.

However, in human-robot interaction, no assumptions about the existence of such boundaries have been made. The focus has been pointed on identifying those factors that influence interaction distance. Interaction distance can be influenced by factors like user age or gender, pet ownership, crowdedness in the environment or available space, as shown by van Oosterhout and Visser (2008); Takayama and Pantofaru (2009). However, their analyses did not include users' perceptions about the behavior or features of the robot.

Proxemics measurement were included, hoping to find some correlations between interaction distance and the factors treated in the questionnaire. Participant behavior was analyzed also from measuring the distances between the face of the robot and the face of the user and between the face of the robot and the hand of the user[5].

As introduced in Subsection 3.3.4, proxemics analysis were done by gathering data from video recorded during the interaction sessions (Figure 3.6 shows a sample frame). The user sat in front of the robot at a distance of ca. 90 *cm*. The interaction was recorded with a standard camera (resolution 640×480) placed at ca. 2 meters perpendicularly to the axis robot-user. Beside the table where the robot was standing there was a scale drawn on a whiteboard for the visual

[3]High Cronbach's alpha values are those greater than 0.5, which specify that the used set of variables are good for defining a certain concept.

[4]For recoding, the variable was flipped: $1 = 5$, $2 = 4$, $3 = 3$, $4 = 2$, $5 = 1$.

[5]When interacting with the robot, participants did not use two hands at the same time. Almost all of them performed movements only with one arm, or at least they alternated between left and right. Only the movements of the active arm were registered.

estimation (estimated average error: $5cm$) of the distance between the nose of the user and the head of the robot and from the hand of the user and the head of the robot. Videos were annotated manually: every 5 seconds the face-face and face-hand distances were visually estimated by the operator, manually projecting their positions onto a scale drawn on the whiteboard.

Participants were sitting on a chair (they all started at the same distance to the robot), but they were told to feel free to interact in any way they considered more appropriate. However, it happened only in very few cases (only 2 participants) that they stood up. In both the cases, the face-face and face-hand distances were gathered as projected onto the horizontal line parallel to the table.

Results

This subsection presents the quantitative evaluation of the experiments. An earlier experiment uncovered some interesting patterns (Hafner and Schillaci, 2011; Bodiroza et al., 2011). It seemed that if a person holds the object close to the robot's hand, then Nao's pointing will be perceived as a desire to grasp the object. This could indicate that, along with the hypothesis that pointing emerges from grasping, there is also a reverse connection – pointing can be perceived as grasping, if the object is too close to the hand[6]. Furthermore, most of the participants in the preliminary experiment responded that Nao was either likeable or very likeable and that the speed of experiment was good (out of three possible answers: too fast, good and too slow), even though the execution speed was lower than in the current experiment. All participants in the preliminary experiment, except one, had no previous experience with robots.

Figure 3.7 shows the means and the standard deviations of the responses.

First, it was checked whether the distributions of the collected data are normal or not, in order to select the proper statistical tests. For each variable (that is, for each question), the superimposition of the histogram of the data with a normal curve characterized by the mean and the variance of the data was checked. Almost all the histograms did not fit well together with the corresponding normal curves. Thus, the kurtosis and the skewness of the data was measured[7], in order to have a more precise measurement of the normality of the distributions. The distributions of all the variables related to the questionnaire had kurtosis and skewness between -2 and $+2$, while 17 out of 64 distributions related to the variables of the proxemics analysis[8] did not.

Due to the non-normality of such distributions, it seems to be more appropriate applying non-parametric statistical tests for the whole analysis. However, the use of ANOVA on Likert-scale data and without the assumption of normality of the distributions of the data to be analyzed is controversial. In general, researchers claim that only non-parametric statistics should be used on Likert-scale data and when the normality assumption is violated. Vallejo et al. (2010), instead, found that the Repeated Measures ANOVA[9] was robust toward the violation of normality assumption. Simulation results of Schmider et al. (2010) confirm also this observation, since they found in their Monte Carlo study that the empirical Type I and Type II errors in ANOVA were not affected by the violation of assumptions.

[6]The used robot platform had no movable fingers and it was therefore unable to grasp an object.

[7]In general, when kurtosis and skewness are between -2 and $+2$, the data is not too far away from a normal distribution. When that is not the case, corrections (like Box-Cox transformations) can be applied to the data in order to apply the tests that have assumptions of normality.

[8]For each of the four behaviors performed by the robot, two variables were created for the average value and variance of the distance between the face of the Nao and the nose of the participant for the following cases: during the first 15 seconds of the interaction, between the 15th second and the 45th second of the interaction, and during the last 15 seconds of the interaction (in total 6 variables). The same variables were created for analyzing the distance between the face of the Nao and the user's hand closest to the robot.

[9]Repeated measures ANOVA compare the average score for a single group of subjects at multiple time periods (observations).

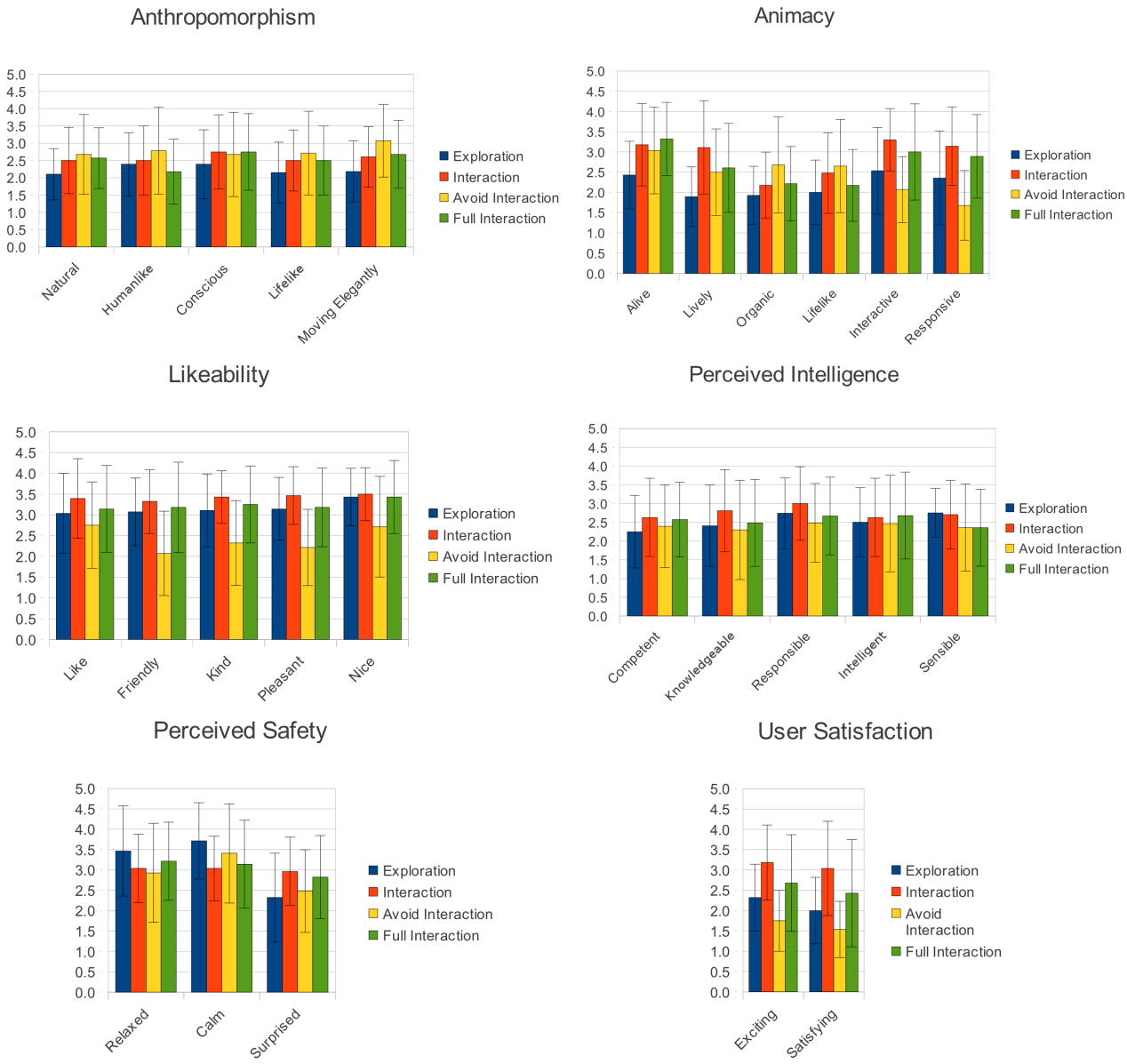


Figure 3.7: Results of the experiment from the Godspeed questionnaire

Correlations

A Spearman's Rank Order correlation[10] was run to determine the relationship between perceived factors and between them and average human-robot distances. Each run was done for each experimental session (*exploration*, *interaction*, *interaction avoidance* and *full interaction*).

Tables 3.1 and 3.2 show some of the most relevant correlations. In addition to the data shown in the tables, it has to be noted that in the *exploration* test there was a strong, positive correlation between almost all the anthropomorphism variables and the perceived intelligence attributes related to competence and knowledge; in *interaction*, the higher the likeability of the robot, the higher the variance of face-face distance during all the interaction tests ($r = 0.805$, $P = 0.029$, $N = 7$); in *full interaction*, perceived intelligence was found to be positively correlated with almost all the other variables (except those related to perceived safety) with $r > 0.5$ and almost always significant at the 0.01 level.

Repeated Measures ANOVA

Because the participants of the four different observations were the same in each group, the Repeated measures ANOVA test was adopted (post-hoc test using Bonferroni correction) for the analysis of variances. Also known as within-subjects ANOVA test, repeated measures ANOVA is the equivalent of the one-way ANOVA but for related, not independent groups. The test was performed on all the dependent variables[11] .

Post-hoc tests revealed that the four different behaviors performed by the robot have not changed significantly the participants' perception of the anthropomorphic attributes related to naturalness, humanlikeness, consciousness and artificiality. Table 3.3 shows the statistically significant results of repeated measures ANOVA on the questionnaire variables.

Proxemics variables contain a high number of missing values. In order to perform repeated measures ANOVA on those variables, missing values had to be replaced with multiple imputation ($n = 20$). New samples were created, where proxemics information was inferred using the questionnaire variables as predictors[12].

Table 3.4 shows the statistically significant results of the repeated measures ANOVA on the proxemics variables.

Latent Growth Curve Model

A latent growth curve model was also used to assess the change in user perception over the four behaviors. This model uses a structural equation to estimate two latent variables, the slope and

[10]Spearman's correlation coefficient is non-parametric, looks at ranked (coded) variables (without looking at the data directly) and does not have the normality assumption on the distributions, thus it can be used for skewed or ordinal variables. The correlation was evaluated with 2-tailed test of significance. Missing values were excluded with cases pairwise.

[11]Mauchly's test has been used as statistical test for validating repeated measures ANOVA. It tests the sphericity, which is related to the equality of the variances of the differences between levels of the repeated measures factor. Sphericity, an assumption of repeated measures ANOVA, requires that the variances for each set of difference scores are equal. Sphericity can not be assumed when the significance level of the Mauchly's test is < 0.05. Violations of sphericity assumption can invalidate the analysis conclusions, but corrections can be applied to alter the degrees of freedom in order to produce a more accurate significance value, like the Greenhouse-Geisser correction. When the significance level of the Greenhouse-Geisser estimate is < 0.05, statistical significant differences revealed by post-hoc test can be elicited from the pairwise comparisons between the observations. Repeated Measures ANOVA does not tell where the differences between groups lie. When repeated measures ANOVA is statistically significant (both with sphericity assumption not violated or with Greenhouse-Geisser correction), post-hoc tests with multiple comparisons can highlight exactly where these differences occur.

[12]For multiple imputation, all the available variables that can predict the values of missing data should be included.

intercept, to assess the average linear change across the measurements, where the individual measurements are the indicators of the latents[13]. The estimated population distribution of the linear change (or growth) trajectory, denoted by the slope and the intercept of a linear function, are derived from this structural equation model. The estimator selected for the procedure was a Bayesian estimator with non-informative priors[14]. All calculations were produced with Mplus 6.11.

The estimated slopes for many of the items were almost all positive, with also positive credibility intervals, meaning that there is a significant positive trend in the average score from the first observation (*exploration*) to the last observation (*full interaction*)[15].

3.4 Discussion and Conclusions

It was shown that by using a combination of bottom-up and top-down processes of attention and the ego-sphere (as explained in Subsection 3.3.2) as a short-term memory representation, it is possible to have an interaction between the robot and a person, even on robots with limited computational capacities such as the Nao. The combination of motion, face and object detection has led to simple interactions between a person and the Nao.

With the current implementation, the Nao was able to process approximately 5-6 frames per second. Future optimizations could improve the speed of execution, but the first results of a qualitative survey show that the speed of experiment is perceived as good even at this state.

Despite the small sample size of the data collected during the experiments (especially regarding the proxemics analysis), the outcomes suggested many elements and features that need to be carefully taken into account when developing attentive mechanisms for intuitive robot behavior.

Godspeed Questionnaire

The adoption of the Godspeed questionnaire allowed us to test its qualities. Questionnaires are important tools in measuring user perceptions and the Godspeed provided us with a good instrument for measuring the quality of the implemented robot behaviors. Its authors noted that comparing different robots and their settings by means of the same measurement index will help roboticists in making design decisions. The indices of the Godspeed questionnaire have been tested as measures of human-like characters (Ho and MacDorman, 2010). The results indicated significant and strong correlations among some relevant indices and new indices have been proposed. This matches the comments of most of the participants of the experiments which complained about the similarity between many questions and about some high-level attributes which were difficult to assign to the robot. The problem with the recorded variable and the previous notes suggest to not adopt the original version of the Godspeed questionnaire for further experiments, but rather its revisited version.

To the best of our knowledge, no other study on attentional mechanisms for robots has adopted the Godspeed questionnaire as a metric. However, Ham et al. (2011) studied the combined and individual contribution of gestures and gazing to the persuasiveness of a story-telling robot measuring user perception with the Godspeed questionnaire. The robots used persuasive gestures (or not) to accompany the persuasive story, and also used gazing (or not) while telling the persuasive

[13]The loadings are constrained to be 1 for the intercept latent and to 0 to 3 (depending of the time of measurement) for the slope latent.

[14]This estimation strategy was appropriate as the more commonly used maximum likelihood estimator often produces biased (or often inestimable) results with such small sample sizes. The Bayesian estimator is more robust to both small samples and violation of distributional assumptions that could emerge from small samples.

[15]Further analysis can be done on piecewise linear growth, for breaking up the curvilinear growth trajectories into separate linear components, thus for analyzing whether there was an increase or a decrease between *exploration* and *interaction*, between (*interaction* to *interaction avoidance*, and so on.

story. Their results indicated that only gazing had a main effect on persuasiveness, while the use of gestures did not. Also, the combined effect of gestures and gazing on persuasiveness was greater than the effect of either gestures or gazing alone. This study suggests that adding multiple social cues can have additive persuasive effects, matching what is to be discussed in the next subsection about multi-modal interaction and efficient feedback systems.

Correlations

Correlation analysis confirmed the initial expectations and suggested directions for improvement of robot attentional mechanisms. Positive correlations between anthropomorphic attributes and perceived intelligence confirmed that a robot with human-like appearance can increase the level of its perceived intelligence. However, an excessive human-like appearance can entail the interacting person having too high expectations about the robot's cognitive capabilities, which can provoke disappointment whenever the robot does not fulfill such expectations. The positive correlations between the anthropomorphic attributes and the perceived intelligence reflect a good balancing between Nao's human-like appearance and its implemented cognitive capabilities. Confirming this hypothesis, most of the participants did not try to communicate vocally with the robot, suggesting that they were not expecting this interaction modality due to the absence of a mouth in the robot's face and due to any other robot's verbal capability.

Positive correlations between the robot's interactiveness and user excitement and perception of lifelikeness and intelligence (see Table 3.1, correlations between Animacy: interactive and Perceived Intelligence: intelligent) suggested also that interactive capabilities emerging from attentional mechanisms can increase the perceived level of intelligence of the robot. Such results confirm also the thought that a robot has to be highly interactive for being perceived as a highly intelligent agent, and it has to be responsive for increasing user satisfaction.

A relevant contribution in the user satisfaction is given by the robot's responsiveness and interactivity and it can be increased by improving its feedback system. A well designed feedbacks system could reduce the consequences of some robot limitations. In the experiments, participants experienced issues related to the limited field of view of the Nao (58°). It is plausible that humans expect of humanoids to have approximately matching characteristics, such as the field of view, or two eyes for vision.[16] During the experiments, participants, without being aware of that, were often waving to the robot or handing over the object out of the robot's field of view causing no reactions to it. This resulted in affecting the perception of the robot's responsiveness and interactiveness. A little foresight in the feedback system could have probably reduced this effect, like changing the color of head LEDs, or emitting sounds, whenever the robot detected something.

Multi-modal interaction (through arm or head movements) increased the level of interactiveness perceived by participants, as suggested by the correlations between Animacy and Interactive, as well as several other variables (see Table 3.1) which during *interaction* were higher than when the robot performed other behaviors[17]. The consideration of Ham et al. (2011) about combining gestures and gazing for increasing the persuasiveness and the likeability of the robot matches the consideration about multi-modal interaction. Elegance in movement positively correlating with user satisfaction suggests that the robot might perform smooth and natural movements in order to increase the quality of the interaction.

A trustworthy and lifelike robot can be better accepted as a companion or as a co-worker, where close interaction is needed, as suggested by the negative correlation between lifelikeness and face-face average distance recorded during *interaction* ($r = 0.805$, $P = 0.029$, $N = 7$).

[16] Video cameras are not located in the positions of the eyes on the Nao, which leads to unmet expectations.
[17] During *interaction*, the robot performed arm and head movements during the whole session.

Repeated Measures ANOVA

Repeated measures ANOVA results showed that the aliveness of the robot during *exploration* scored lower than during *interaction* and *full interaction*, again supporting the initial expectation that multi-modal interaction increases the expressiveness of the robot behaviors (in *exploration*, the robot performed only head movements). Again, more expressive movements or a better designed feedback system could have increased the level of perceived animacy, likeability and user satisfaction.

The less the interaction was perceived as satisfactory, the more often and the more frenetically the participants moved their hand. Repeated measures ANOVA confirmed that the variance of face-hand distance is higher during *interaction avoidance* (the least satisfactory robot behavior for the users) than during the other behaviors. It is also interesting to note how successful the *interaction avoidance* behavior was, by which the robot did cause frustration to the users, according to its *motivation* of avoiding the interaction. Several participants commented this behavior assigning mental states to the robot, like *shyness* and *angriness*.

A saliency based attentional model combined with a robot ego-sphere and implemented it on a humanoid robot was introduced. Human-robot interaction experiments, in which this model was used, presented that different attentional behavior of the robot has a strong influence on the interaction as experienced by the human.

It was shown that it is possible to have an ongoing interaction between the robot and a person, even on robots with limited computational capacities such as the Nao. Techniques used are a combination of bottom-up and top-down processes of attention and an ego-sphere as a short-term memory representation in combination with motion, face and object detection.

The adopted questionnaires were useful for correlating perceived physical and behavioral robot features with proxemics data. Some trends were noticed suggesting that some of the perceived variables could influence the distances during interaction.

Through the discussion of the results in the previous section, those characteristics that need to be emphasized and those skills that have to be taken into account (like providing enough feedback during the interaction) when implementing attentive mechanisms in robots were identified.

These experiments represent a step in a good direction toward reaching joint attention between a human and a robot. It was shown that basic attention manipulation is possible, even with simple robot platforms, such as the Nao, and that participants will assign different characteristics to it based on its behavior.

Following abbreviations are used throughout tabular results: Ani – Animacy, Ant – Anthropomorphism, Lik – Likeability, PI – Perceived Intelligence, PS – Perceived Safety, US – User Satisfaction, avg – average, var – variance, FF – distance between the face of the robot and the face of the user, FH – distance between the face of the robot and the closest hand of the user, all – considering the whole duration of the test (60 seconds).

Table 3.1: Most relevant correlations (part 1).

		Exploration			Interaction			Inter. Avoidance			Full Interaction		
Variables correlated		R	p	N	R	p	N	R	p	N	R	p	N
Ant: humanlike	Ani: alive	0.581	0.001	28	0.654	0.000	28	0.665	0.000	28	0.546	0.003	28
Ant: humanlike	Ani: interactive	0.513	0.005	28	0.605	0.001	27				0.504	0.006	28
Ant: humanlike	PI: knowledgeable	0.416	0.031	28	0.562	0.003	26	0.571	0.002	27	0.606	0.001	27
Ant: humanlike	PI: competent	0.476	0.011	28	0.677	0.000	27	0.623	0.000	28	0.564	0.002	28
Ant: humanlike	PI: intelligent				0.559	0.002	27	0.573	0.001	28	**0.713**	0.000	28
Ant: natural	PI: knowledgeable	0.498	0.008	27	0.557	0.003	26				0.553	0.003	27
Ant: natural	PI: competent	0.565	0.002	28	0.612	0.001	27				0.572	0.001	28
Ant: moving elegantly	PI: knowledgeable	0.697	0.000	26	0.399	0.044	26	0.654	0.000	27	0.422	0.028	27
Ant: moving elegantly	PI: competent	**0.694**	0.000	27	0.483	0.011	27	0.542	0.003	28	0.454	0.015	28
Ant: moving elegantly	Lik: friendly							-0.500	0.007	28			
Ant: lifelike	var FH ($15s-45s$)										**-0.786**	0.012	9
Ani: lifelike	Lik: friendly				0.663	0.001	23	-0.425	0.043	23			
Ani: interactive	Ant: lifelike	**0.673**	0.012	13	0.660	0.000	27				0.556	0.002	28
Ani: interactive	Lik: friendly	0.398	0.036	28	0.451	0.018	27				0.655	0.000	28
Ani: interactive	PI: intelligent	**0.462**	0.013	28	**0.705**	0.000	26	**0.403**	0.033	28	**0.619**	0.000	28
Ani: interactive	US: exciting	**0.710**	0.000	28	0.551	0.004	26				**0.706**	0.000	28
Ani: interactive	US: satisfying	0.470	0.012	28	0.687	0.000	26				0.725	0.000	28
Ani: responsive	avg FF all ($60s$)	0.633	0.037	11									
US: satisfying interaction	Ant: moving elegantly	0.505	0.007	27	0.482	0.011	27				0.390	0.040	28
US: satisfying interaction	Ant: lifelike	0.576	0.002	26	**0.653**	0.000	27				0.516	0.005	28
US: satisfying interaction	Ani: responsive	**0.696**	0.000	28	**0.722**	0.000	27				**0.673**	0.000	28
PS: quiescent	avg FF (last $15s$)							-0.879	0.009	7			
PS: quiescent	avg FH (last $15s$)							-0.805	0.029	7			

Table 3.2: Most relevant correlations (part 2).

		Exploration			Interaction			Inter. Avoidance			Full Interaction		
Variables correlated		R	p	N	R	p	N	R	p	N	R	p	N
var FH ($0s - 15s$)	PS: quiescent										-0.670	0.048	9
var FH ($60s$)	Lik: friendly				0.802	0.030	7				-0.673	0.047	9
var FH ($60s$)	Lik: kind										**-0.738**	0.023	9
var FH ($60s$)	Lik: pleasant										**-0.829**	0.006	9
var FH ($60s$)	US: satisfying interaction										**-0.738**	0.023	9
var FF dist. ($15s - 45s$)	Lik: friendly				0.809	0.028	7						
avg FF ($60s$)	US: exciting										0.709	0.032	9
avg FF ($60s$)	PI: intelligent										0.729	0.026	9

Table 3.3: Statistically significant results of repeated measures ANOVA on the questionnaire variables. Cases with sphericity assumption violated were corrected with Greenhouse-Geisser method. The table shows the statistically significant pairwise comparisons (illustrating the changes in means from an observation to another), taken from the post-hoc test with Bonferroni correction.

Variable	Sphericity assumed	From observ.	To observ.	Mean difference	Std. error	Significance
Ant: moving elegantly	no	1	3	-0.889	0.252	0.010
		2	3	-0.481	0.154	0.026
Ani: alive	yes	1	2	-0.75	0.203	0.006
		1	4	-0.893	0.165	0.000
Ani: lively	yes	1	2	-1.222	0.284	0.001
		1	4	-0.741	0.224	0.016
Ani: organic	no	1	2	-0.750	0.239	0.025
		2	3	-0.500	0.159	0.024
Ani: interactive	yes	1	2	-0.815	0.251	0.019
		2	3	1.222	0.202	0.000
		3	4	-1.000	0.233	0.001
Ani: responsive	yes	1	2	-0.786	0.259	0.032
		2	3	1.464	0.260	0.000
		3	4	-1.214	0.243	0.000
Lik: friendly	no	1	3	1.000	0.230	0.001
		2	3	1.250	0.270	0.001
		3	4	-1.107	0.274	0.002
Lik: kind	yes	1	3	0.786	0.243	0.019
		2	3	1.107	0.248	0.001
		3	4	-0.929	0.224	0.002
Lik: pleasant	no	1	3	0.929	0.185	0.000
		2	3	1.250	0.222	0.000
		3	4	-0.964	0.238	0.002
Lik: nice	no	1	3	0.714	0.198	0.008
		2	3	0.786	0.249	0.023
PS: quiescent	yes	1	2	0.593	0.194	0.031
		2	3	-0.481	0.154	0.026
US: exciting	no	1	2	-0.852	0.218	0.004
		2	3	1.407	0.234	0.000
		2	4	0.444	0.154	0.047
		3	4	-0.963	0.285	0.014
US: satisfying	yes	1	2	-1.037	0.196	0.000
		2	3	1.519	0.222	0.000
		3	4	-0.963	0.229	0.002

Table 3.4: Statistically significant results of repeated measures ANOVA on the proxemics variables. Cases with sphericity assumption violated were corrected with Greenhouse-Geisser method. The table shows the statistically significant pairwise comparisons (illustrating the changes in means from an observation to another), taken from the post-hoc test with Bonferroni correction. Missing values were replaced with multiple imputations. The new dataset contained 560 samples.

Variable	Sphericity assumed	From observ.	To observ.	Mean difference	Std. error	Significance
avg FF all	no	1	2	**13.675**	0.419	0.000
		1	3	11.876	0.302	0.000
		1	4	9.734	0.355	0.000
		2	3	**-1.799**	0.360	0.000
		2	4	**-3.941**	0.436	0.000
		3	4	-2.142	0.280	0.000
var FF all	no	1	2	-6.218	2.058	0.016
		1	3	-82.552	3.014	0.000
		1	4	**14.558**	1.914	0.000
		2	3	-76.334	2.361	0.000
		2	4	20.776	1.633	0.000
		3	4	97.110	2.862	0.000
avg FH all	no	1	2	14.767	0.544	0.000
		1	3	**18.439**	0.547	0,000
		1	4	**21.423**	0.539	0.000
		2	3	3.672	0.294	0.000
		2	4	6.656	0.314	0.000
		3	4	2.985	0.307	0.000
var FH all	no	1	2	**20.337**	3.861	0.000
		1	3	**-217.131**	7.246	0.000
		2	3	**-237.468**	6.718	0.000
		2	4	-28.076	5.602	0.000
		3	4	**209.392**	6.904	0.000

Chapter 4

Gesture Vocabularies for Human-Robot Interaction

4.1 Overview

This chapter presents three experiments on gesture vocabularies. In relation to Figure 1.1, it covers the topics from defining use-case scenarios up to obtaining gesture vocabularies and their analysis, as well as improvement of robot gestures through application of interactive genetic algorithms, as presented in Figure 4.1. This is the first part that concerns the topic of use of gestures in human-robot interaction. In the next chapter, the second part will be discussed.

Section 4.2 presents background work in the topic of design of human gesture vocabularies and discusses robot gesturing and how robot morphology might affect it. Section 4.3 provides a definition of gesture vocabularies.

Section 4.4 introduces an example use-case scenario describing interaction between a customer and a robot waiter, based on interpersonal interactions observed in bars or restaurants. Based on this, two experiments are performed to develop gesture vocabularies for a human and a robot. Both experiments are performed as user surveys. In the first experiment on human gesture vocabularies the participants were asked to produce a gesture they associate with particular actions, which were afterwards coded by a person, in order to discover the mappings between gestures and their associated actions. In the second survey the participants were asked to rate different qualities of gestures produced by a robot in a simulation, as well as rank them.

An approach to refining of robot gestures based on interactive genetic algorithms is presented in Section 4.5. The approach uses a human rater as a fitness function for rating robot gestures based on their aesthetic qualities. The ratings are used as fitness values for standard procedures of producing offspring in classical genetic algorithms.

The work presented in this chapter is based on work presented in Bodiroža et al. (2012).

4.2 Background

Stern et al. (2008) approached development of human gesture vocabularies as a multi-objective optimization problem, taking in consideration three performance measures: intuitiveness, stress/comfort, and accuracy. The first measure presents how intuitively a gesture represents its associated command. The second measure relates to the effort required to perform a particular gesture or the stress. The last measure reflects the ability of a gesture recognition system to accurately recognize

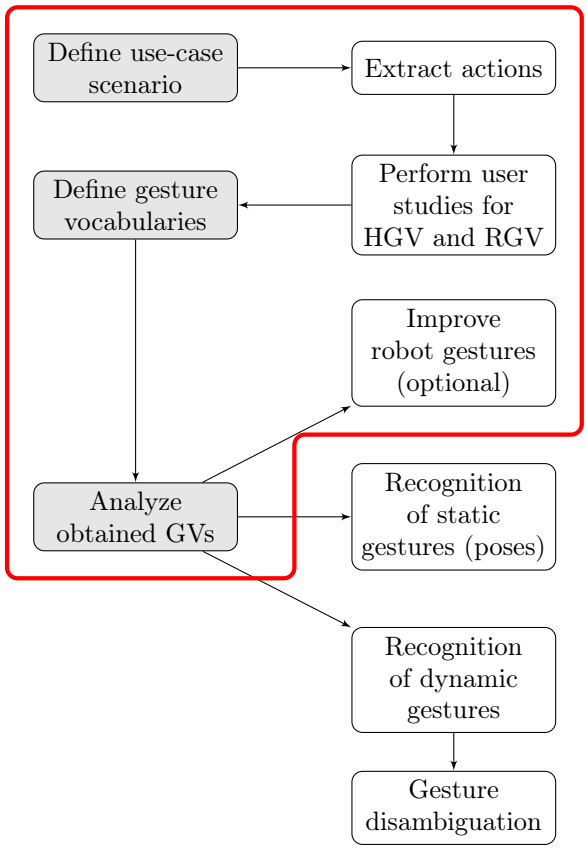


Figure 4.1: Research topics covered in Chapter 4

gestures from the defined gesture vocabulary. In addition, it is mentioned that a good gesture vocabulary should be intuitive, physically easy to perform, easy to recognize by a recognition system, easy to learn and to remember, as a consequence of intuitiveness and effortlessness. In this work, development of human gesture vocabularies will be approached from another side, that is instead of having persons choose gestures from a predefined set, they will be asked to freely produce gestures.

Interaction modalities, such as visual, auditory and gestural, have been employed to provide feedback to a listener. Rea et al. (2012) used an iRobot Roomba augmented with colored lights to communicate the mood of the others present in the room. Mazzei et al. (2012), as well as Bazo et al. (2010) developed systems for generating facial expressions to convey a range of emotional states. Salem et al. (2010) worked on run-time production of related co-speech gestures. Some of those communication channels are present in interpersonal interaction, such as speech and gestures, while other are not, such as use of chirping sounds and indicator lights. Hegel et al. (2011) worked on developing a set of guidelines for design of those feedback signals. However, in order for interaction to be intuitive to a person, gestures performed by a robot must be intuitive. Production of human and robot gestures is affected by multiple aspects. Namely, human gestures are affected by the quality of gesture recognition algorithms, and production of robot gestures is affected by robot's morphology. These aspects are reviewed in the rest of the section.

Gesture recognition and gesture production are primary research directions in the area of gestures in human-robot interaction. The former considers the task of classifying the perceived motion of the hand and arm into one of the gesture classes. While the state of the art has improved over time, some issues are still commonly present, such as occlusion, both self-occlusion and occlusion by another object, and recognition when the orientation between the person and the robot differs from the one during the learning. Some of these methods are also capable of learning new gestures, when the classification is unsuccessful, in which case these gestures are assigned to new clusters. On the other hand, perception of robot gestures has not been well investigated, and there is a need for work in this area.

Gesture production is another important area. In here, the goal is synthesis and execution of gestures, which are intuitive to a human person observing the robot. The research can be divided into following main subareas. On one side, there is research on generation of co-speech, iconic gestures (Ng-Thow-Hing et al., 2010; Wachsmuth and Kopp, 2002). This is usually achieved through analysis of videos of narratives, where a person is describing some story and observing which co-speech gestures appear, with which words and what is the particular timing between a vocal and a gesture utterance. A gesture production system is then trained to learn these connections and autonomously produce co-speech gestures with speech, taking in consideration the particular timing of gestures. Since each trained model is tied to a particular person, gesturing of different persons can be modeled and emulated. Another area concerns the trajectory interpolation of gestures. A gesture can be represented as a sequence of 3D points in space describing the position of an end-effector at particular points in time. The easiest way to define the complete gesture trajectory from keyframes is to perform linear interpolation of points between two consecutive keyframes. However, this often results in rigid, linear and chopped movements. Other interpolation methods can be applied in order to generate curved motion, which is perceived as more natural and human-like. These are usually based on B-splines. Keyframes represent geometric knots, through which a spline curve passes. Parametric knots are then computed based on the geometric knots, which are used to compute the interpolated curved trajectory.

A robot can learn gestures through imitation or demonstration. These gestures can also be adapted and improve using interactive genetic algorithms. The problem can arise in the difference between morphologies of robot's and person's arm and hand. Ende et al. (2011) reported that a particular subset of their gesture vocabulary had higher human recognition rate when they were performed by a human or a humanoid robot, than when they were performed by an industrial manipulator attached to the ceiling.

Actuation of the manipulators can be achieved using different methods. Usual are serially connected, serially actuated joint segments, pull cable-driven segments and pneumatically actuated segments. Serially connected, serially actuated joint segments have an actuator for each degree of freedom. In a humanoid arm, these joints are revolute joints, performing a rotational motion. An example can be seen in robots such as Aldebaran Nao and Sony Aibo. Pull cable-driven joint segments are actuated by tensile forces in a pull cable around the joint. An example can be seen in Ecce robot, which is designed to mimic human motion (Marques et al., 2010; Potkonjak et al., 2011).

Another important aspect for gesture production is the morphology of the hand and the arm. A lack of hand dexterity in a robot can prevent it from performing gestures, where important information is encoded in a hand pose (e.g., gestures typically used for counting, or the "victory" gesture, where the index and the middle fingers resemble a "V" shape). Based on the observations on how people gesture, a set of gestures is defined as a gesture vocabulary for the robot (Bodiroža et al., 2012). Due to the similarities in morphologies of the arm in a human and in a Nao robot, the participants of the experiments were still able to understand and rate well the robot's gesture, as it is shown in the Section 4.4.2.

There has been little research in development of robot gesture vocabularies, with an exception of work by Ende et al. (2011), and it was usually dealing with a small gesture vocabulary for a few actions, where each action was represented with one particular gesture. The proposed approach takes into account that a robot can use more gestures to represent one action and that these gestures should be drawn from a pool of alternative gestures, which is defined for each action. The goal of this is to increase the quality of the human-robot interaction, through making the robot seem more human-like, in the way that it can use different gesture alternatives to represent a single action.

4.3 Defining Gesture Vocabularies

Human-human interaction commonly relies on numerous communication channels, such as speech, facial expressions, and gestures, in order to have a redundant and reliable transfer of information, and so that the message can be reconstructed even when one part of it is lost due to the noise (e.g., a person talking and gesturing in a loud environment). Similarly, the communication in human-robot interaction should be multi-modal to achieve more robust and accurate communication. However, the accuracy of various recognition algorithms, including gesture recognition, is still not sufficiently high.

Chapter 3 presented the importance of joint attention for successful human-robot interaction. However, there are other contributing factors. One of them are efficient means for information transfer, whether it is communicating intentions, or giving a command. Gestures are featured prominently in interpersonal interaction. A person gestures consciously, e.g., when pointing out directions, but they can also occur unconsciously during the interaction within a group of people. Two persons interacting do not need to see each other for gestures to appear. Gesturing is observed in such cases when a person is talking over a phone, or when a person is blind (Iverson and Goldin-Meadow, 1998). Therefore, they could be used as an intuitive communication method in human-robot communication. However, meaning of a particular gesture is highly context-dependent and is influenced by various factors, such as cultural background. In order to alleviate this issue, gesture vocabularies can be used. They represent a bidirectional, commonly many-to-many mapping between a set of meanings and a set of gestures, as defined below.

Definition 4.1 *Given a set of* ***meanings***

$$M = M_a \cup M_i \cup M_m$$

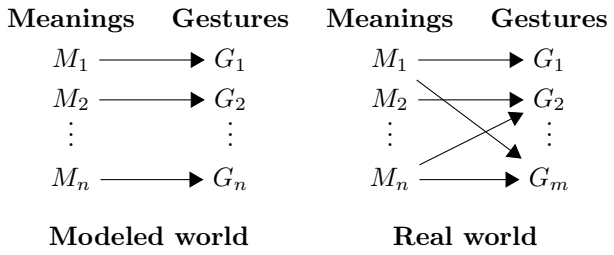


Figure 4.2: Example of a designed gesture vocabulary for a modeled world, and a gesture vocabulary seen in the real world

consisting of sets of ***actions****,* $M_a = \{m_{a1}, \ldots, m_{ai}\}$*,* ***intentions****,* $M_i = \{m_{i1}, \ldots, m_{ij}\}$*, and* ***messages****,* $M_m = \{m_{m1}, \ldots, m_{mk}\}$*, and a set of* ***gestures***

$$G = G_{arms} \cup G_{head} \cup \ldots$$

consisting of all body gestures, $G_i = \{g_{i1}, \ldots, g_{im}\}$*, where* i *denotes a particular body part or multiple body parts, a* ***gesture vocabulary*** *is represented with multi-valued functions* v_1 *and* v_2*, such that*

$$v_1 \colon M \to G, v_2 \colon G \to M$$

mapping ***meanings*** *and* ***gestures****.*

When the development of gesture vocabularies is not given enough forethought, e.g., when they are defined by the application designer, it can lead to a decrease in the intuitiveness of interaction. On the other side, by carefully designing gesture vocabularies, building on the results of the analysis of interpersonal interaction, the issue of decreased intuitiveness could be partially prevented.

This chapter is partially based on a paper by Bodiroža et al. (2012), underlain by the work of Stern et al. (2006).

4.4 Case Study: Gesture Vocabularies for a Robot Waiter Scenario

A robot waiter scenario is used as an example scenario to display the procedure for obtaining two gesture vocabularies. The scenario considers usual interaction occurring between a customer and a waiter. Two gesture vocabularies, that is sets of gesture-meaning mappings are obtained, one for a robot waiter and one for a person interacting with it. First, dialogs which occur in natural customer-waiter interaction were obtained through observation of those interactions. These were used as a basis for obtaining a set of discrete, or atomic meanings, which are initiated either by a customer or by a waiter. A set of gestures, corresponding to the meanings, is then defined through a user survey or by observing human participants in case of human gesturing, or by designing robot gestures in case of robot gesturing. Currently, most of the gesture vocabularies are usually hand-designed. While being practical, it excludes potential users, which will later interact with the robot, and it can therefore influence the intuitiveness of the system to the user.

In the current scenario, one of the expectations is to see prevalent use of deictic (i.e., pointing gestures) and symbolic gestures in developed gesture vocabularies, as primary goal is to indicate the intention of the user. Most of the gestures are expected to be codified. These gestures, contrary to

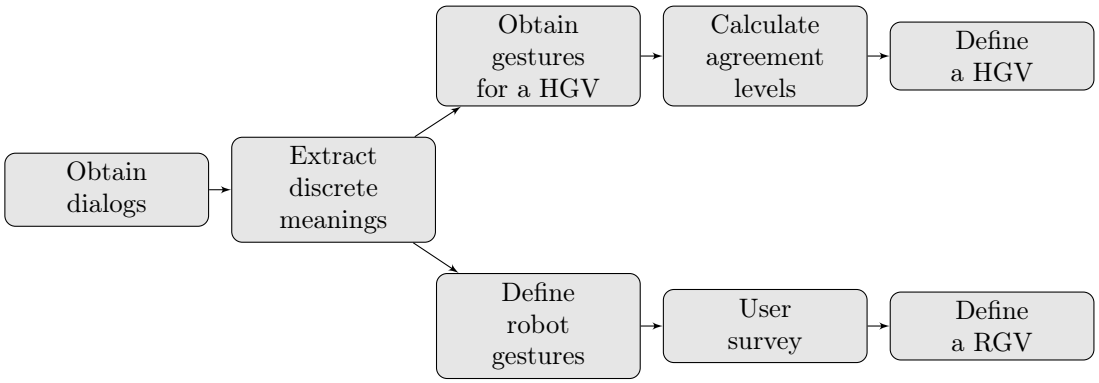


Figure 4.3: Procedure for obtaining human and robot gesture vocabularies

a creative gesture, is the one which has a fixed representation in our own gesture vocabularies, with mapping between the sensorimotor signals (motor commands and perception) and the semantic information it carries (Poggi, 2002). A preliminary experiment showed that participants indeed used prevalently deictic and symbolic gestures (Bodiroža et al., 2012). Participants used deictic gestures to specify an object of their attention. The gesture refers to an object of interest and manipulates the attention of the recipient, e.g. a robot waiter, and the transferred message can be implied in certain situations. For example, pointing to an empty glass can indicate a request for another glass of the same drink.

Stern et al. define three performance measurements for gesture usability: intuitiveness, comfort and recognition rate (Stern et al., 2006,0). The first two measures are human-related, while the last if algorithm-related. In author's recent work preliminary results showed that most of the participants will choose the same gesture, or two gestures will emerge among the others (Bodiroža et al., 2012).

A usual and easy approach for definition of human gesture vocabularies is to predefine the gestures which need to be recognized by the algorithm, as well as their meanings, such as in the work of Ende et al. (2011) and Mai et al. (2011).

However, it is common to observe many-to-many mapping of gestures and their meanings in a human gesture vocabulary. Allowing many-to-many mappings increases intuitiveness as it partially relaxes the limitations placed on the user, due to the occurrence of alternative gestures for single actions. However, it also introduces the need for analysis of contextual information in the gesture recognition process (e.g., in the case of a pointing gesture, to which object the user is pointing).

A flow diagram representing a method of obtaining human and robot gesture vocabularies is presented in Figure 4.3.

4.4.1 Human Gesture Vocabulary

A human gesture vocabulary can be obtained through different approaches. A naïve approach would be to hard-code a set of gestures which can be used to transfer meanings to a robot. In this case, the mappings between gestures and their respective meanings are defined by the system designer. However, this approach might prove not to be intuitive for an everyday user, as gestures are not universal (Kendon, 2004) and gesture models, that is the way a particular person associates gestures with their meanings, differ between people (Bergmann et al., 2010).

A procedure for obtaining an intuitive human gesture vocabulary is proposed here. It uses a human-in-the-loop model, where a resulting human gesture vocabulary highly depends on charac-

teristics of a group of experiment participants. By using this model, a possible bias of the system designer is removed from the vocabulary. In summary, the proposed procedure is as follows:

- Through analysis of a scenario, define a set of actions that are initiated by a person (e.g., asking for a menu in a restaurant),
- Conduct a survey, in which each action is presented to participants, and asking them to perform a gesture they associate with a respective action, in case they can relate a gesture to that action,
- Find gestures with high agreement levels, that is a number which represents percentage of users that chose that gesture as a representative for a particular action, and
- Define a human gesture vocabulary, by creating relations between actions and their representative gestures, that is those that had high agreement levels.

Expected results are to have one or two gestures per each action, which are ranked higher than others. Additional questions in the survey (such as speed and precision) can be used to modify the selected gesture (e.g. if the speed is not adequate or if the trajectory is not good enough). A proposal to improve the results using interactive genetic algorithms is introduced in Section 4.6.1.

Experimental Setup and Procedures

A method for selection of a human gesture vocabulary using a survey was developed (Bodiroža et al., 2012).

Prerequisite for the determination of a human gesture vocabulary is a scenario, which describes the nature of the interaction (e.g. the interaction between a customer and a waiter, or within rescue teams). The scenario contains a list of actions that are carried out in these interactions (e.g. order a drink or clean the table in the robot waiter scenario). Furthermore, a target user group needs to be defined and survey participants need to represent well this group. It is suggested that gesture vocabularies are not culturally universal (Graham and Argyle, 1975; Morris, 1979; Walker and Nazmi, 1979; Kendon, 2004). Therefore, a gesture vocabulary obtained in one culture might not be adequate for another.

Contrary to the usual approach of introducing restrictions on which kind of gestures can be used, the following method does not put these constraints on participants. The expected results is to get gestures which can be classified in the following categories: one gesture representing the whole service, two gestures – one representing the action itself, and the other the object, or pointing gestures, which indicate the currently attended object.

Participants performing gestures were recorded with Microsoft Kinect, a RGB-D camera sensor capable of recording both color and depth images. A sequence of recorded color images was converted into a video. All videos were then coded by a person. OpenNI framework was used for skeletal tracking of participants from depth images. Skeletal tracking provides a 3D location of participants in Kinect's frame of reference and this data can be used for testing of a gesture recognition system. The Kinect was placed below a display with a diagonal of 50 inches at a distance of $2m$ from the participant.

Each participant was first introduced to the experimental procedure, including the means and the goals of the experiment, and was asked to sign a consent form, approved by the Ethical Committee of the Ben-Gurion University of the Negev. They were sitting down during the experiment. Each session took around 15 minutes. The actions are displayed to the participant in a sequential order. Each action is represented with a sentence, that would be used in order to carry out the service (e.g. "I would like to order a drink." in the case of the service order a drink). Each sentence has marked keywords, which represent the action (i.e. a verb) and the object (i.e. a noun).

The robot platform is the Nao from Aldebaran, a humanoid robot around 57*cm* tall. For the experiment, only the degrees of freedom in the arms and the neck were used – 4 degrees of freedom in each arm and 2 degrees of freedom in the neck. This robot was chosen due to the movable arms and humanoid shape. Moreover, its small child-resembling size could reduce users' expectations, thus increasing the positive evaluation of the interactions. The gestures' motions were implemented using MotionNet implemented in C++ using the framework of the Nao Team Humboldt[1]. All the gestures were performed and recorded with a video-camera.

The participants were 8 Industrial Engineering students ($mean\ age = 28.75, \sigma = 3.77$) from Ben-Gurion University of the Negev.

Results

The results are presented in Table 4.1. The preselected meanings, based on the described scenario, were "call waiter" (1), "order beer" (2), "cancel beer" (3), "order this" (4), "cancel this" (5), "ask for a suggestion" (6), "clean table" (7), "take away glass" (8), "bring bill" (9), "take away bill" (10) and "where is the toilet" (11). Participants used following gestures (agreement level $AL > 25\%$): pointing (1), writing on an imaginary piece of paper (2), index finger wave (3), hand wave "no" (4), sliding gesture for canceling (5), circular movement of hand over a surface (6), circular movement of finger over a surface (7), handling the object (8), raised hand (9), hand wave (10), no gesture (11). Tasks were call waiter (1), order beer (2), cancel beer (3), order this (4), cancel this (5), ask for a suggestion (6), clean table (7), take away glass (8), bring bill (9), take away bill (10) and where is the toilette (11).

Table 4.1: Results of the human gesture vocabulary experiment. Columns are meanings, rows are gestures, values represent percentage of people that used a specific gesture for a specific task.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	0	**62**	0	**100**	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	**75**	0	0
3	0	0	**37**	0	**25**	0	0	0	0	0	0
4	0	0	**25**	0	0	0	0	0	0	0	0
5	0	0	0	0	**37**	0	0	0	0	0	0
6	0	0	0	0	0	**25**	**100**	0	0	0	0
7	0	0	0	0	0	**25**	0	0	0	0	0
8	0	0	0	0	0	0	0	**75**	0	**87**	0
9	**37**	0	0	0	0	0	0	0	0	0	0
10	**50**	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	**25**	0	0	0	0	**62**

The tasks with high agreement levels were easy to characterize either with action, deictic, or iconic gestures. Tasks with low agreement levels could not be easily characterized. This yields one guideline for the task analysis – chosen tasks should be easy to depict with gestures. It was expected that participants will use certain gestures that are fairly regular for certain tasks (e.g., an action gesture simulating drinking for ordering a beer), such as with tasks *Order a beer* and *Order this*. But, this was not the case, since a fairly high number of participants used pointing (to a menu) for both tasks. Some gestures were used for more tasks. Unlike the approach where one limits gestures for certain tasks, such as in Mai et al. (2011), here is proposed that a robot should consider other contextual information, e.g., if the person is holding a menu or if there is a spill on the table. Thus, two-way connections between gestures and tasks must be created. As all agreement levels are above 25% it is easy to extract the G-T pairs from Table 4.1 that constitute such a unique complex GV.

[1]Motion generating tool, a part of the Nao Team Humboldt's framework for the Aldebaran Nao.

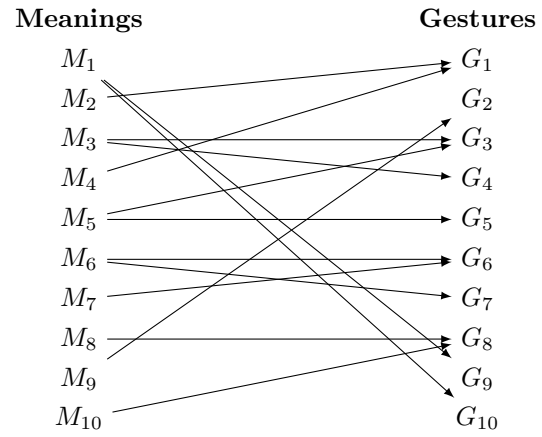


Figure 4.4: Obtained human gesture vocabulary, with excluded gesture 11 ("no gesture") and meaning 11 ("Where is the toilet?")

4.4.2 Robot Gesture Vocabulary

Human-human interaction uses different modalities to transfer a message. These modalities can be combined to increase readability and robustness of the communication. Gestures are one communication channel and they often co-occur with speech. In order to be more readable, a humanoid robot should produce gestures during its interaction with a human, when necessary.

Previously, researchers have focused on work related to the lifelike gesture production (Ng-Thow-Hing et al., 2010; Wachsmuth and Kopp, 2002), which usually focused on how to adapt the robot gestures' trajectories and co-expressive timing to resemble human gestures. On the other hand, there has been some work in how well participants understand certain robot gestures (Ende et al., 2011). However, to the best of author's knowledge, there has been no attempt of devising a robot gesture vocabulary by letting participants rank gesture-action mappings with multiple proposed gestures for a particular action.

Objective

The goal of this work is to develop a method for defining robot gesture vocabularies for a particular scenario. The procedure is similar to the procedure for obtaining a human gesture vocabulary and is as follows:

- Analyze the scenario and define a set of actions that can be initiated and performed by a robot,
- Develop few robot gestures per each action, where the gestures are observed from the interpersonal interaction,
- Conduct a survey asking participants to rate the gestures for each action, in terms of how well do they fit a particular action,
- Conduct statistical analysis to find gestures which are ranked better than others,
- Define a robot gesture vocabulary, containing gesture-action mappings.

Expected results are to have one or two gestures per each action, which outrank other proposed gesture for the specified action. Additional questions in the survey (such as speed and precision) can be used to modify the selected gesture (e.g. if the speed is not adequate or if the trajectory is not good enough). A proposal to improve the results using interactive genetic algorithms is introduced in the Future work section.

Robot Platform Aldebaran Nao

The robot platform used for the experiment is the RoboCup version of Aldebaran Nao, a humanoid robot around 57cm tall with 21 degrees of freedom. For the purpose of the experiment, only the degrees of freedom in the arms and the neck were used – 4 degrees of freedom in each arm and 2 degrees of freedom in the neck. This robot was chosen because it has movable arms and a humanoid shape, which can make it easier on the human user to understand gestures (see results of study by Ende et al. (2011)). Moreover, its small child-resembling size could reduce users' expectations, thus increasing the positive evaluation of the interactions. The gesture trajectories were implemented using MotionNet implemented in C++ using the framework of the Nao Team Humboldt[2]. All the gestures were recorded with a standard video camera.

User experiments

In the experiment to develop a robot gesture vocabulary participants were asked to rate gestures for eight different actions, listed in Table 4.2. There were two or three alternative gestures for each action. These gestures were performed by Aldebaran Nao, a humanoid robot, and were video-recorded.

The experimental procedure was as follows. The participants first watched all video recordings of representative gestures of one action in a randomized order. They were asked to rate the following three characteristics for each alternative gesture: overall impression (scale $1 \dots 9$), speed (scale $-4 \dots 4$) and precision (scale $1 \dots 9$)[3]. They were also asked to rank the displayed gestures from the best to the worst, where worst rated would get assigned a rating of 1, while each better would have a rating incremented by 1.

The participants were 21 Israeli students (13 male, 8 female, $mean\ age = 25.38$, $\sigma = 2.27$) from Ben-Gurion University of the Negev.

A part of the designed gestures, representing actions presented both in human and robot vocabularies, were the same as those obtained in the human gesture vocabulary experiment. The intention was to test whether these gestures, that are also present in a human gesture vocabulary, will achieve higher ratings. Table 4.2 presents the actions and their alternative gestures[4].

In the following text, the gestures will be numbered, instead of being named (e.g., "gesture 4.1", instead of "Pointing to the cup as a sign to take it away").

Analysis

The collected data was analyzed using repeated measures ANOVA (independent variables were "speed", "precision" and "overall impression", dependent variables were gestures). Additionally, Spearman's rank was determined with bivariate correlation to determine statistically significant correlations between independent variables.

[2]Motion generating tool, a part of the Nao Team Humboldt's framework for the Aldebaran Nao.

[3]Precision was defined as how smooth was the gesture trajectory or how precise was the pointing.

[4]The RoboCup version of the Nao robot has no movable fingers, therefore pointing is performed with the hand and not with the finger

Table 4.2: Selected actions and their representative gestures

1. “Hi, my name is Robowa.”	1. Pointing with the hand to the name tag on the chest.
	2. Putting hands close together in front of self.
	3. Waving with the hand.
2. “Did you call me?”	1. Pointing with the right hand to the name tag on the chest.
	2. Right arm a bit stretched, as if it is offering something.
3. “Would you like to order?”	1. Waving with the hand up-down in front of the menu.
	2. Pointing with the right arm to the menu.
	3. Both arms stretched and open, with the palms up.
4. “Can I take away this cup?”	1. Pointing with the right hand to the cup.
	2. Raising right hand to the chest and waving away towards the cup.
5. “Would you like me to suggest a special drink?”	1. Pointing with the right arm to the menu.
	2. Circling with the hand in front of the menu.
	3. Waving with the hand up-down in front of the menu.
6. “Would you like me to bring you a napkin?”	1. Rubbing the mouth with the hand.
	2. Right arm a bit stretched, as if it is offering something.
7. “Do you want me to clean the table?”	1. Circling with the hand over the table.
	2. Raising right hand to the chest and waving away towards the table.
8. “Do you want the bill?”	1. Making a checkmark sign with the right hand in the air.
	2. Making a writing movement with the left hand over the right hand.
	3. Making a rectangular sign with both hands, indicating the shape of the bill.

To calculate the correlations, variable "speed" was recoded in the following manner: $-4 = 0$, $-3 = 1$, $-2 = 2$, $-1 = 3$, $0 = 4$, $1 = 3$, $2 = 2$, $3 = 1$, $4 = 0$; and variable "ranking" was recoded in the following manner: $1 = 3$, $2 = 2$, $3 = 1$.

Results

All correlation results report Spearman's rho values (Tables 4.3–4.10). The number of data samples in all cases was $N = 21$. Statistically significant correlations are marked with bold font, where one asterisk denotes significance at the 0.05 level, and two asterisks denote significance at the 0.01 level.

Table 4.3: Correlations for gesture alternatives of action 1.

Gesture alternative 1		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	.393	**.441***
	Sig. (2-tailed)	.	.078	.045
Speed	Corr. Coefficient	.393	1.000	.305
	Sig. (2-tailed)	.078	.	.179
Precision	Corr. Coefficient	**.441***	.305	1.000
	Sig. (2-tailed)	.045	.179	.
Gesture alternative 2		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	**.436***	**.730****
	Sig. (2-tailed)	.	.048	.000
Speed	Corr. Coefficient	**.436***	1.000	.339
	Sig. (2-tailed)	.048	.	.133
Precision	Corr. Coefficient	**.730****	.339	1.000
	Sig. (2-tailed)	.000	.133	.

Table 4.4: Correlations for gesture alternatives of action 2.

Gesture alternative 1		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	.288	**.519***
	Sig. (2-tailed)	.	.206	.016
Speed	Corr. Coefficient	.288	1.000	**.624****
	Sig. (2-tailed)	.206	.	.002
Precision	Corr. Coefficient	**.519***	**.624****	1.000
	Sig. (2-tailed)	.016	.002	.
Gesture alternative 2		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	**.578****	**.783****
	Sig. (2-tailed)	.	.006	.000
Speed	Corr. Coefficient	**.578****	1.000	.394
	Sig. (2-tailed)	.006	.	.077
Precision	Corr. Coefficient	**.783****	.394	1.000
	Sig. (2-tailed)	.000	.077	.

Results indicate significant correlation between "precision" and "overall impression" for most gestures. "Speed" does not seem to be highly correlated with any of the other two variables. There were in total 20 gestures, in 18 of them, "precision" and "overall impression" were correlated, in 7 of them "speed" and "precision" were correlated, in 4 of them "speed" and "overall impression" were correlated.

It was assumed that there will be a significant correlation between "speed" and "overall impression" in most cases. However, results indicate that the speed of the gesture do not affect much its overall impression.

Furthermore, there was correlation between "overall impression" and the ranking of the gesture for most of the alternatives. Additionally, for some smaller subset, there was also correlation between the "precision" and the ranking of the alternatives. No significant correlations were found between the "speed" variable and rankings.

Table 4.5: Correlations for gesture alternatives of action 3.

Gesture alternative 1		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	-.024	**.469***
	Sig. (2-tailed)	.	.916	.032
Speed	Corr. Coefficient	-.024	1.000	.417
	Sig. (2-tailed)	.916	.	.060
Precision	Corr. Coefficient	**.469***	.417	1.000
	Sig. (2-tailed)	.032	.060	.
Gesture alternative 2		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	.141	**.701****
	Sig. (2-tailed)	.	.541	.000
Speed	Corr. Coefficient	.141	1.000	**-.018**
	Sig. (2-tailed)	.541	.	.938
Precision	Corr. Coefficient	**.701****	-.018	1.000
	Sig. (2-tailed)	.000	.938	.
Gesture alternative 3		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	**.696****	**.736****
	Sig. (2-tailed)	.	.000	.000
Speed	Corr. Coefficient	**.696****	1.000	**.567****
	Sig. (2-tailed)	.000	.	.007
Precision	Corr. Coefficient	**.736****	**.567****	1.000
	Sig. (2-tailed)	.000	.007	.

Table 4.6: Correlations for gesture alternatives of action 4.

Gesture alternative 1		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	.086	**.661****
	Sig. (2-tailed)	.	.710	.001
Speed	Corr. Coefficient	.086	1.000	.187
	Sig. (2-tailed)	.710	.	.418
Precision	Corr. Coefficient	**.661****	.187	1.000
	Sig. (2-tailed)	.001	.418	.
Gesture alternative 2		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	**.446***	**.612****
	Sig. (2-tailed)	.	.043	.003
Speed	Corr. Coefficient	**.446***	1.000	.403
	Sig. (2-tailed)	.043	.	.070
Precision	Corr. Coefficient	**.612****	.403	1.000
	Sig. (2-tailed)	.003	.070	.

Table 4.7: Correlations for gesture alternatives of action 5.

Gesture alternative 1		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	.336	**631****
	Sig. (2-tailed)	.	.136	.002
Speed	Corr. Coefficient	.336	1.000	.315
	Sig. (2-tailed)	.136	.	.165
Precision	Corr. Coefficient	**.631****	.315	1.000
	Sig. (2-tailed)	.002	.165	.
Gesture alternative 2		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	.203	**.575****
	Sig. (2-tailed)	.	.378	.006
Speed	Corr. Coefficient	.203	1.000	.102
	Sig. (2-tailed)	.378	.	.661
Precision	Corr. Coefficient	**.575****	.102	1.000
	Sig. (2-tailed)	.006	.661	.

Developed Robot Gesture Vocabulary

Based on the result of the statistical analysis, following gestures were chosen for a robot gesture vocabulary:

- Action "Hi, my name is Robowa": G1.1 Pointing with the hand to the name tag on the chest.
- Action "Did you call me?": G2.1 Pointing with the right hand to the name tag on the chest.

Table 4.8: Correlations for gesture alternatives of action 6.

Gesture alternative 1		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	.352	**.703****
	Sig. (2-tailed)	.	.117	.000
Speed	Corr. Coefficient	.352	1.000	**.494****
	Sig. (2-tailed)	.117	.	.023
Precision	Corr. Coefficient	**.703****	**.494****	1.000
	Sig. (2-tailed)	.000	.023	.
Gesture alternative 2		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	.202	**.487***
	Sig. (2-tailed)	.	.381	.025
Speed	Corr. Coefficient	.202	1.000	**.473***
	Sig. (2-tailed)	.381	.	.030
Precision	Corr. Coefficient	**.487***	**.473***	1.000
	Sig. (2-tailed)	.025	.030	.

Table 4.9: Correlations for gesture alternatives of action 7.

Gesture alternative 1		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	.398	**.533***
	Sig. (2-tailed)	.	.074	.013
Speed	Corr. Coefficient	.398	1.000	.420
	Sig. (2-tailed)	.074	.	.058
Precision	Corr. Coefficient	**.533***	.420	1.000
	Sig. (2-tailed)	.013	.058	.
Gesture alternative 2		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	.387	**.620****
	Sig. (2-tailed)	.	.083	.003
Speed	Corr. Coefficient	.387	1.000	**.473***
	Sig. (2-tailed)	.083	.	.030
Precision	Corr. Coefficient	**.620****	**.473***	1.000
	Sig. (2-tailed)	.003	.030	.

Table 4.10: Correlations for gesture alternatives of action 8.

Gesture alternative 1		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	.424	**.552****
	Sig. (2-tailed)	.	.055	.009
Speed	Corr. Coefficient	.424	1.000	**.735****
	Sig. (2-tailed)	.055	.	.000
Precision	Corr. Coefficient	**.552****	**.735****	1.000
	Sig. (2-tailed)	.009	.000	.
Gesture alternative 2		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	.176	**.785****
	Sig. (2-tailed)	.	.445	.000
Speed	Corr. Coefficient	.176	1.000	.210
	Sig. (2-tailed)	.445	.	.362
Precision	Corr. Coefficient	**.785****	.210	1.000
	Sig. (2-tailed)	.000	.362	.
Gesture alternative 3		Overall impression	Speed	Precision
Overall impression	Corr. Coefficient	1.000	.210	**.582****
	Sig. (2-tailed)	.	.362	.006
Speed	Corr. Coefficient	.210	1.000	**.446***
	Sig. (2-tailed)	.362	.	.043
Precision	Corr. Coefficient	**.582****	**.446***	1.000
	Sig. (2-tailed)	.006	.043	.

- Action "Would you like to order?": G3.1 Waving with the hand up-down in front of the menu.
- Action "Can I take away this cup?": G4.1 Pointing with the right hand to the cup.
- Action "Would you like me to bring you a napkin?": G6.1 Rubbing the mouth with the hand.
- Action "Do you want me to clean the table?": G7.1 Circling with the hand over the table.

For following actions, no single gestures are determined:

- Action "Would you like me to suggest a special drink?"
- Action "Do you want the bill?"

In this section, detailed results for all actions are presented.

Action "Hi, my name is Robowa"

This action represents a robot introducing itself. Gesture 1.1 ranked better than both gestures 1.2 and 1.3 ($r_1 = 2.65$, $r_2 = 1.55$, $r_3 = 1.8$, $p < 0.01$) . The mean overall impression of the gesture 1.1 was higher than of the gesture 1.2 ($p < 0.01$).

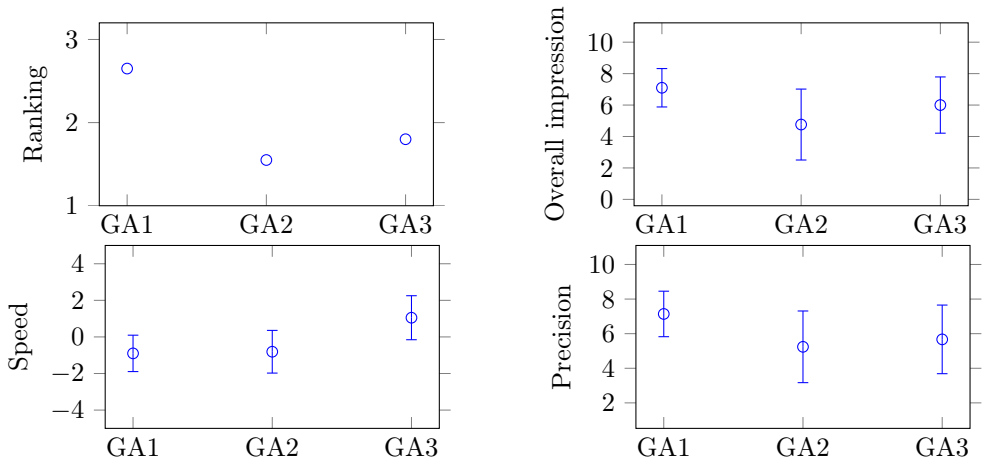


Figure 4.5: Rankings and mean overall impressions of gesture alternatives for action 1

Table 4.11: Means and standard deviations (in brackets) for overall impression, precision and speed of gesture alternatives of action 1.

	Overall impression	Speed	Precision
Gesture alternative 1	7.10 (1.221)	-0.90 (0.995)	7.14 (1.315)
Gesture alternative 2	4.76 (2.256)	-0.81 (1.167)	5.24 (2.071)
Gesture alternative 3	6.00 (1.789)	1.05 (1.203)	5.67 (1.983)

Action "Did you call me?"

The action is executed when the robot assumes that a guest called it. Gesture 2.1 ranked better than gesture 2.2 ($r_1 = 2.85$, $r_2 = 2.15$, $p < 0.01$). The mean OI of the gesture 2.1 was higher than of the gesture 2.2 ($p < 0.01$).

Action "Would you like to order?"

This gesture is performed when robot would like to serve a guest. Gestures 3.1 and 3.2 ranked higher than gesture 3.3 ($r_1 = 2.6$, $r_2 = 2.1$, $r_3 = 1.3$, $p < 0.01$). The mean OI of the gesture 3.1 was higher than of the gesture 3.3 ($p < 0.01$).

Action "Can I take away this cup?"

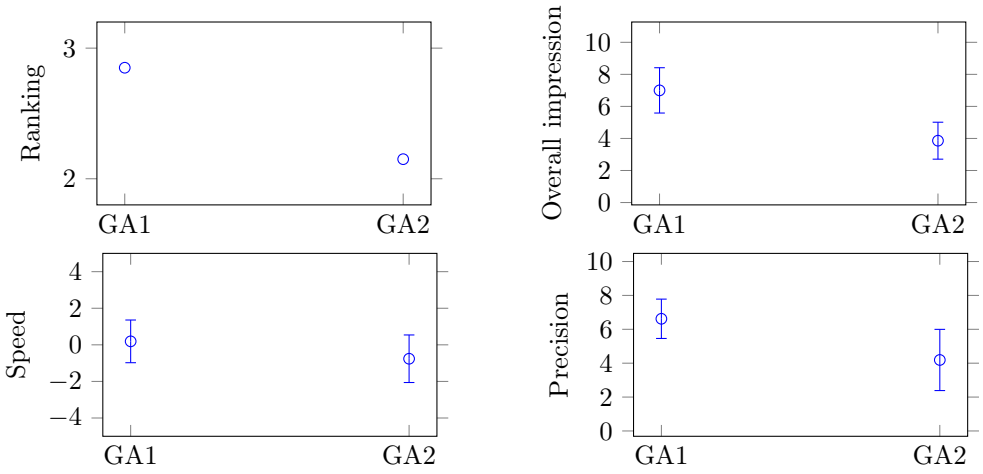


Figure 4.6: Rankings and mean overall impressions of gesture alternatives for action 2

Table 4.12: Means and standard deviations (in brackets) for overall impression, precision and speed of gesture alternatives of action 2.

	Overall impression	Speed	Precision
Gesture alternative 1	7.00 (1.414)	0.19 (1.167)	6.62 (1.161)
Gesture alternative 2	3.86 (1.153)	-0.76 (1.300)	4.19 (1.806)

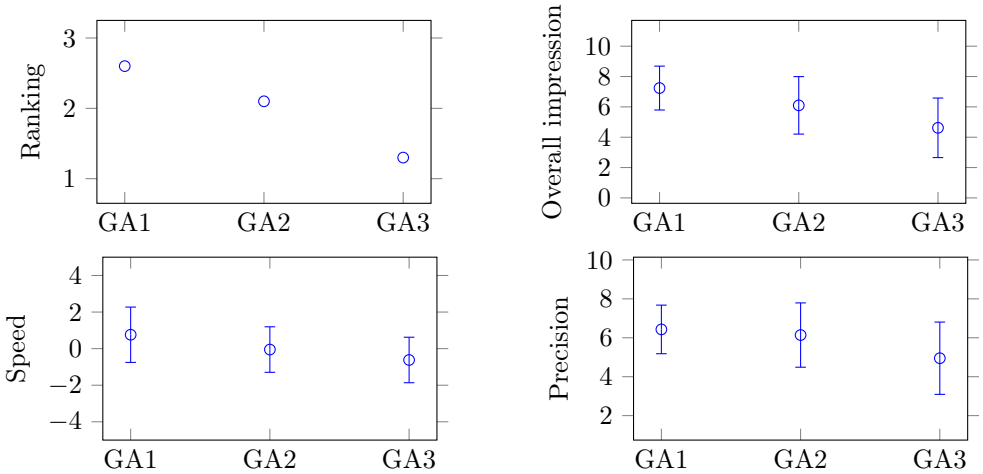


Figure 4.7: Rankings and mean overall impressions of gesture alternatives for action 3

This gesture is performed when robot would like to take away an empty cup. Gesture 4.1 ranked higher than the gesture 4.2 ($r_1 = 2.9$, $r_2 = 2.1$, $p < 0.01$) The mean OI of the gesture 4.1 was higher than of the gesture 4.2 ($p < 0.01$).

Action "Would you like me to suggest a special drink?"

This gesture is executed when the robot would like to make a suggestion (e.g. if the guest is

Table 4.13: Means and standard deviations (in brackets) for overall impression, precision and speed of gesture alternatives of action 3.

	Overall impression	Speed	Precision
Gesture alternative 1	7.24 (1.446)	0.76 (1.513)	6.43 (1.248)
Gesture alternative 2	6.10 (1.895)	-0.05 (1.244)	6.14 (1.652)
Gesture alternative 3	4.62 (1.962)	-0.62 (1.244)	4.95 (1.857)

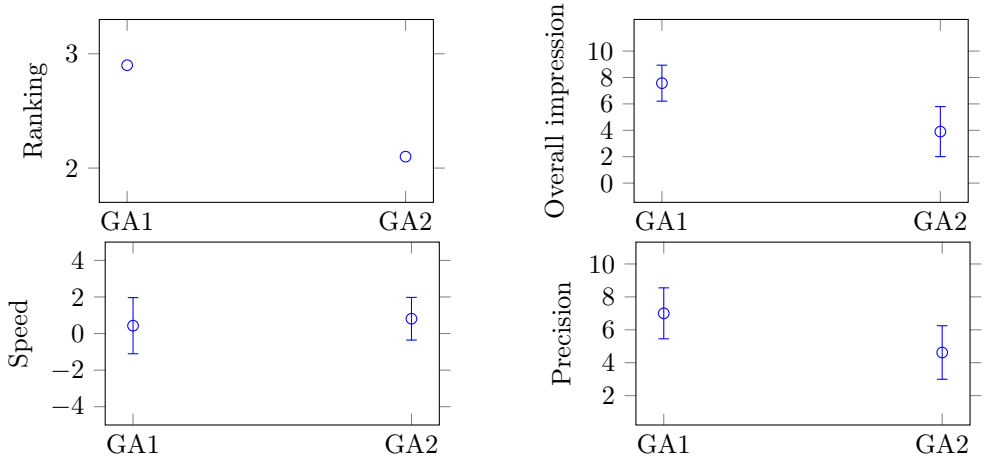


Figure 4.8: Rankings and mean overall impressions of gesture alternatives for action 4

Table 4.14: Means and standard deviations (in brackets) for overall impression, precision and speed of gesture alternatives of action 4.

	Overall impression	Speed	Precision
Gesture alternative 1	7.57 (1.363)	0.43 (1.535)	7.00 (1.549)
Gesture alternative 2	3.90 (1.895)	0.81 (1.167)	4.62 (1.627)

indecisive). There were no significant differences between these gestures.

Table 4.15: Means and standard deviations (in brackets) for overall impression, precision and speed of gesture alternatives of action 5.

	Overall impression	Speed	Precision
Gesture alternative 1	6.19 (1.965)	-0.33 (1.317)	5.90 (1.895)
Gesture alternative 2	5.05 (1.717)	1.00 (1.483)	5.29 (1.821)
Gesture alternative 3	6.43 (1.248)	0.00 (1.049)	5.86 (1.352)

Action "Would you like me to bring you a napkin?"

In case the robot notices the customer needs a napkin, it will perform this gesture. Gesture 6.1 ranked higher than the gesture 6.2 ($r_1 = 2.8$, $r_2 = 2.2$, $p < 0.01$). The mean OI of the gesture 6.1 was higher than of the gesture 6.2 ($p < 0.01$).

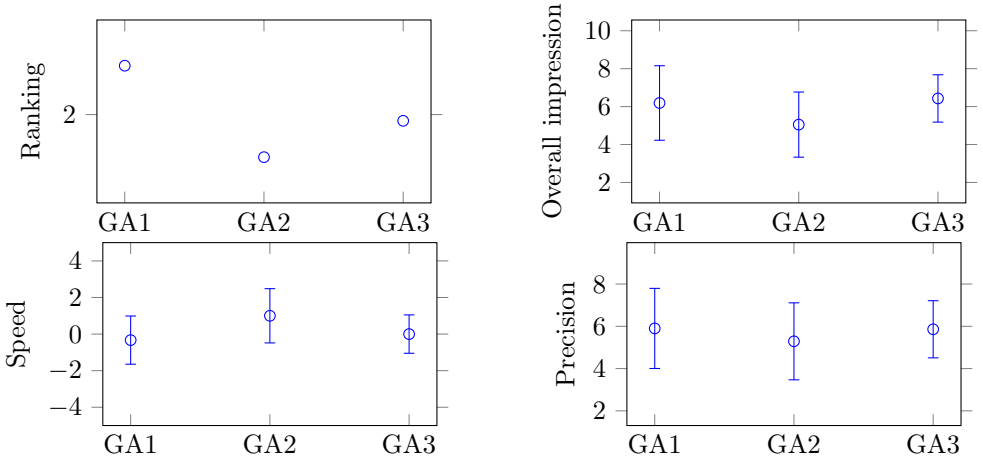


Figure 4.9: Rankings and mean overall impressions of gesture alternatives for action 5

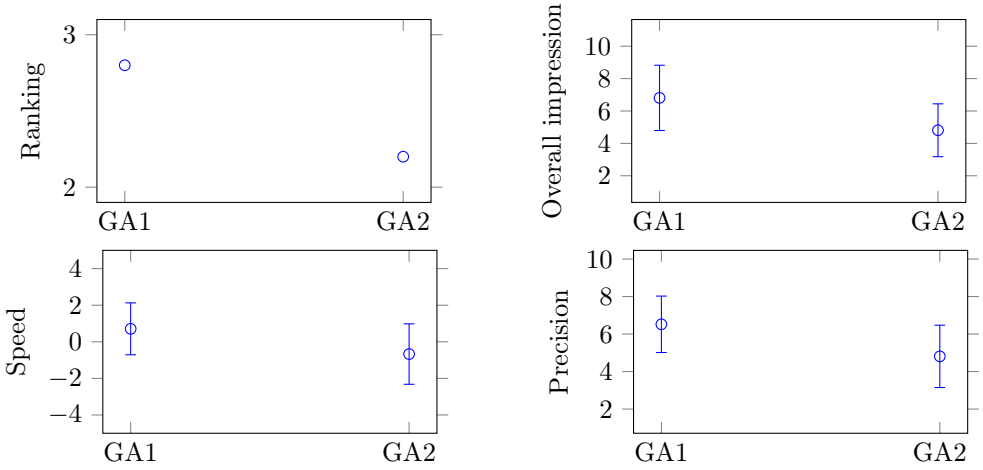


Figure 4.10: Rankings and mean overall impressions of gesture alternatives for action 6

Table 4.16: Means and standard deviations (in brackets) for overall impression, precision and speed of gesture alternatives of action 6.

	Overall impression	Speed	Precision
Gesture alternative 1	6.81 (2.015)	0.71 (1.419)	6.52 (1.504)
Gesture alternative 2	4.81 (1.632)	-0.67 (1.653)	4.81 (1.662)

Action "Do you want me to clean the table?"

When the customer makes a spill, the robot can be proactive and perform a gesture, suggesting it will clean the table. Gesture 7.1 scored higher than the gesture 7.2 ($r_1 = 2.8$, $r_2 = 2.2$, $p < 0.01$). The mean OI of the gesture 7.1 was higher than of the gesture 7.2 ($p < 0.01$).

Action "Do you want the bill?"

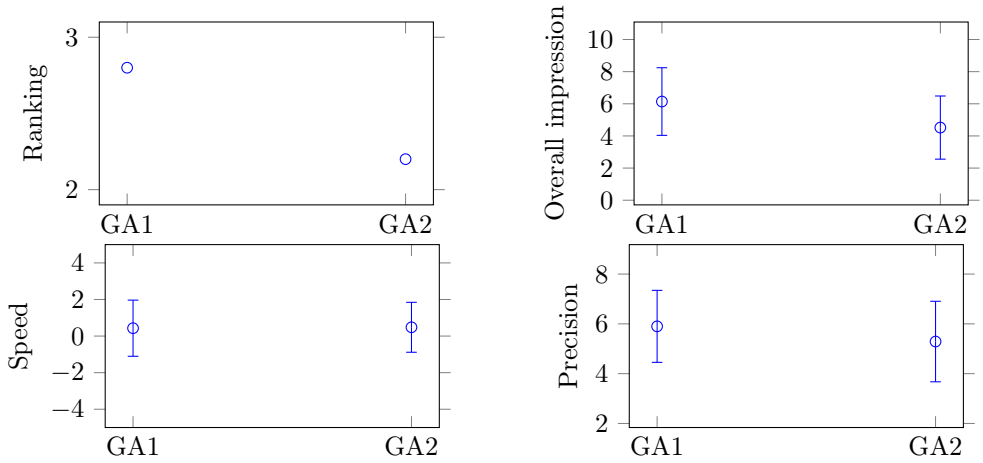


Figure 4.11: Rankings and mean overall impressions of gesture alternatives for action 7

Table 4.17: Means and standard deviations (in brackets) for overall impression, precision and speed of gesture alternatives of action 7.

	Overall impression	Speed	Precision
Gesture alternative 1	6.14 (2.104)	0.43 (1.535)	5.90 (1.446)
Gesture alternative 2	4.52 (1.965)	0.48 (1.365)	5.29 (1.617)

There were no significant differences between these gestures.

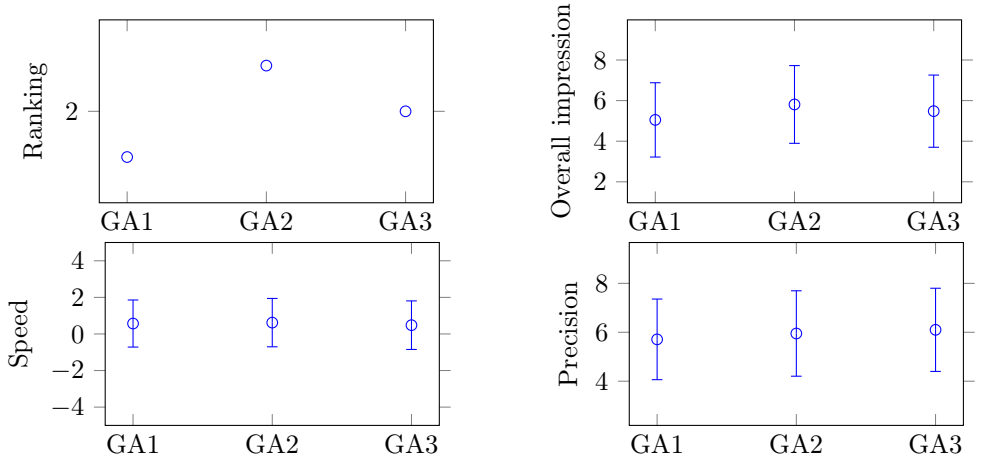


Figure 4.12: Rankings and mean overall impressions of gesture alternatives for action 8

Discussion

There were statistically significant differences in rankings of gestures in 6 out of 8 actions. The results lead to the set of gestures for the Aldebaran Nao in a robot waiter scenario presented in

Table 4.18: Means and standard deviations (in brackets) for overall impression, precision and speed of gesture alternatives of action 8.

	Overall impression	Speed	Precision
Gesture alternative 1	5.05 (1.830)	0.57 (1.287)	5.71 (1.648)
Gesture alternative 2	5.81 (1.914)	0.62 (1.322)	5.95 (1.746)
Gesture alternative 3	5.48 (1.778)	0.48 (1.327)	6.10 (1.700)

Table 4.19: Resulting robot gesture vocabulary

1. "Hi, my name is Robowa."	1. Pointing with the hand to the name tag on the chest.
2. "Did you call me?"	1. Pointing with the right hand to the name tag on the chest.
3. "Would you like to order?"	1. Waving with the hand up-down in front of the menu. 2. Pointing with the right arm to the menu.
4. "Can I take away this cup?"	1. Pointing with the right hand to the cup.
5. "Would you like me to suggest a special drink?"	No significant difference in rankings.
6. "Would you like me to bring you a napkin?"	1. Rubbing the mouth with the hand.
7. "Do you want me to clean the table?"	1. Circling with the hand over the table.
8. "Do you want the bill?"	No significant difference in rankings.

Table 4.19.

In the case of actions for suggesting a special drink and asking if the customer wants a bill there were no statistically significant differences in rankings. A possible explanation for why is it so in the case of the former action is that it might be too specific, and that participants would not associate any particular gesture with this action in interpersonal interaction too. For the latter action, the possible reason might be that proposed gestures cannot be easily understood when performed on this particular robot.

General Remarks

Some general remarks were provided by the participants, who were asked after the experiment to suggest which features would overall improve the understanding of the robot's gestures. Six out of 21 said they would prefer the robot to also include sound feedback in the form of speech. Also, six participants said that human-like hand and fingers, as well as their movements would be preferred to the current state[5]. Four participants said they prefer the introduction of head movements. Four participants noted some gestures should differentiate more, stating that the gestures for two different actions would appear the same to them. The use of interactive genetic algorithms can improve the gestures, through adaptation of robot gestures by employing aesthetic selection, as described in Section 4.5.

[5]Aldebaran Nao RoboCup version was used for the experiment, which has no movable fingers or rotatable wrist.

4.4.3 Comparison of Present Actions in Human and Robot Gesture Vocabularies

Some actions are present in both human and robot gesture vocabularies, namely: ordering a drink, asking for a suggestion, asking to clean the table, asking to take away a glass or a cup, and asking for a bill. For ordering a drink, based on the results of the preliminary human gesture vocabulary experiment (Bodiroža et al., 2012), 62% of the participants used pointing in case where the sentence was "I would like to order a beer", and 100% chose pointing to the menu when the sentence was "I would like to order this". Although a gesture where the robot waves over the menu is selected here, another alternative is pointing which was also rated relatively high (means of the overall impressions of the gestures 3.1 and 3.2 are $oi_{31} = 7.24$ and $oi_{32} = 6.09$, respectively). When participants wanted to ask the robot for a suggestion, 50% of them used a circular movement of the hand or the finger over the menu. However, in the case of the robot gesture vocabulary, significant differences in the rankings and the means of the overall impressions for this action were not found. In order to ask the robot to clean the table, 100% of the participants opted to use the gesture where they move the hand in a circular way over the table. In the case of the robot gestures, another group of the participants rated the same gesture the highest, only performed by the robot ($oi_{71} = 6.14$). The precision of the gesture wasn't that high ($p_{71} = 5.9$), which probably affected the overall impression (Spearman's rank between these two variables for this gesture is $r = 0.53$, $p < 0.01$). Most of the participants, in particular 75% of them, would handle the object when they wanted the robot to take it away. Only small number of them would perform a gesture which would simulate the action of pushing away the object with the hand. Similarly, the same gesture performed by the robot, G4.2, was rated lower than the gesture in which the robot points to the cup, G4.1 ($oi_{41} = 7.57$, $oi_{42} = 3.9$, $p < 0.01$). The action of a robot pointing to the object can be seen as analogous to a person handling the object. When the participants wanted to ask for the bill, most of them, in particular 75%, would perform the gesture in which they write on an imaginary piece of the paper in the air. However, in the case of the robot performing the gesture, no significant difference was found between the three alternatives, which indicates this gesture should be improved.

4.5 An Evolutionary Approach for Improving Robot Gestures

The results of the experiment for selecting gestures for a robot gesture vocabulary showed that not all actions had highly rated gestures. This outcome led to the proposal of employing an evolutionary approach for improving robot gestures. The approach that was employed was an interactive genetic algorithm. Genetic algorithm represent a class of search algorithm. They are inspired by the principles of natural selection, namely mating, selection and mutation. Following subsection provides a more in-depth description of the foundations of genetic algorithms.

4.5.1 Genetic Algorithms

Beginning with a preliminary survey, Holland (1992) poses a problem of adaptive systems consisting of three founding blocks:

- E, the environment of the system under adaptation
- τ, an adaptation plan applied to the system's structures, with the goal of generating improvement, and

- μ, a way to evaluate the result of the adaptation.

Further on, he defined a general framework $(\mathfrak{J}, \mathcal{E}, \chi)$, along the lines outlined in the preliminary survey, where $\mathfrak{J}$ represents a set of adaptive plans to be compared, $\mathcal{E}$ possible environments (uncertainties) of an adaptive system, and χ a criterion for comparing plans in the set of $\mathfrak{J}$.

An adaptive system within this framework is defined as a set of objects $(\mathbb{A}, \Omega, I, \tau)$, where $\mathbb{A} = \{A_1, A_2, \dots\}$ is the set of attainable structures, the domain of action of the adaptive plan, $\Omega = \{\omega_1, \omega_2, \dots\}$ is the set of operators for modifying structures, with $\omega \in \Omega$ being a function $\omega\colon \mathbb{A} \to \mathbb{P}$, where $\mathbb{P}$ is some set of probability distributions over $\mathbb{A}$, I is the set of possible inputs to the system from the environment, and $\tau\colon I \times \mathbb{A} \to \Omega$ is the adaptive plan which, on the basis of the input and structure at time t, determines what operator is to be applied at the time t.

On the example of genetics, he defined following meaning of objects:

- $\mathbb{A}$, populations of chromosomes represented, for example, by the set of distributions over the set of genotypes $\mathbb{A}_1$,
- Ω, genetic operators such as mutation, crossover, inversion, and so on,
- $\mathfrak{J}$, reproductive plans combining duplication according to fitness with the application of genetic operators; for example, if each operator $\omega_i \in \Omega$ is applied to individuals with a fixed probability p_i, then the set of possible plans can be represented by the set,

$$\{(p_1, \dots, p_i, \dots, p_b) \text{ where } 0 \leqq p_i \leqq 1\}$$

- $\mathcal{E}$, the set of possible fitness functions $\{\mu_e\colon \mathbb{A} \to \mathbb{R}\}$, each perhaps stated as a function of combinations of co-adapted sets,
- χ, comparison of plans according to average fitnesses of the populations produced; for example,

$$\inf_{E \in \mathcal{E}} \inf_{t} \inf_{\tau' \in \mathfrak{J}} \overline{\mu}_E(\tau, t) / \overline{\mu}_E(\tau', t)$$

Adaptive plans applied on the general framework, lead to the genetic plans – adaptive plans using generalized genetic operators. A genetic algorithm can be described as follows:

(a) Set $t = 0$ and initialize a population $\mathbb{B}$ by selecting M structures at random from $\mathbb{A}_1$ to form $\mathbb{B}(0) = \{A_h(0), h = 1, \dots, M\}$.

(b) (i) Observe and store the performances $\{\mu_E(A_h(0)), h = 1, \dots, M\}$. Proceed to step (c).
(ii) Observe the performance of $A'(t)$ and replace $\mu_E(A_{j(t)}(t)$ by $\mu_E(A'(t))$.

(c) Increment t by 1.

(d) Select one structure $A_{i(t)}(t)$ from $\mathbb{B}(t)$ by taking one sample of $\mathbb{B}(t)$ using the probabilities $\text{Prob}(A_h(t)) = \mu_E(A_h(t)) / \sum_{h'=1}^{M} \mu_E(A_{h'}(t)), h = 1, \dots, M$.

(e) Determine the operator $\omega_t \in \Omega$ to be applied to $A_{i(t)}, \omega_t = \rho(A_{i(t)}(t))$, and then use ω_t to determine a new structure $A'(t)$ by taking a sample of $\mathbb{A}_1$ according to the probability distribution $P_t = \omega_t(i(t), A_1(t), \dots, A_M(t)) \in \mathbb{P}_{\wp}$.

(f) Assign probability $1/M$ to each number $1, \dots, M$, select one number $1 \leqq j(t) \leqq M$ accordingly, and replace $A_{j(t)}(t)$ by $A'(t)$. Proceed to step (b.i).

$\mathbb{B}(t)$ represents a population at a time t, consisting of its members, represented by structures $A_i, i = 1, \ldots, M$. By applying genetic operators Ω to a selected member $A_{i(t)}(t)$, which can be seen as a "parent", a new member $A'(t)$ is produced, which can be seen as an "offspring" of its "parent". As each new member $A'(t)$ replaces an older member $A_{j(t)}(t)$, the size of $\mathbb{B}(t)$ remains constant through time. Function ρ determines which genetic operator are appropriate to a particular member of the population. With some operators, such as with a cross-over operator, a second member from the population needs to be selected, e.g., as a "mate" for $A_{i(t)}(t)$, that is:

$$\mathbb{A}(t) = (A_1(t), \ldots, A_M(t), \mu_E(A_1(t)), \ldots, \mu_E(A_M(t)))$$

The state is defined with the population $\mathbb{B}(t)$, as well as with the performances $\mu_E(A_h(t)), h = 1, \ldots, M$ of the structures of the population.

4.5.2 Interactive Genetic Algorithms

Interactive genetic algorithms represent a class of algorithms, in which set $\mathcal{E} = \{\mu_e\}$ is replaced with a human rater, performing the role of a fitness function. This is in particular applied when the selection is aesthetic in its nature. A defining characteristic of the aesthetic selection is that it is hard or impossible to define a function μ_e, which evaluates aesthetic qualities of a population $\mathbb{B}$. On the other hand, the quality of "appealing" comes naturally to humans and in this case, they can be employed to evaluate members of a population, seen as using them instead of explicitly defined fitness function.

An example for aesthetic selection is evolution of biomorphs. Biomorphs were created by Dawkins (2013). They represent a 2-dimensional shape resembling a living organism, and they were used as an example of artificial evolution of shapes in order to illustrate the process of natural evolution. In the book, some particular examples of biomorphs were presented, that looked "interesting" to a human eye. This type of evolution, where a fitness of a member is evaluated based on its aesthetic qualities is a problem that can be approached with interactive genetic algorithms and it was addressed in a work by Smith (1991). The initial population was randomly generated and a participant was asked to rate the presented members of the population. Through use of genetic operators $\omega \in \Omega$, together with the ratings used as fitness values for the selection of mating partners, new members were created, which replaced older members of the population.

A very important characteristic, that specifically differentiates interactive genetic algorithms from standard ones, is that the size of population $\mathbb{B}(t)$ is limited to a low number, as well as the possible number of generations $t = 1, \ldots, N$, given that a human rater cannot be asked to evaluate thousands of members $A_i(t)$ for a large number of generations.

Another experiment was performed by Horowitz (1994), in which he presented a procedure for learning user's preferences for generating musical rhythms.

4.5.3 Evolution of Robot Gestures

Given the results of the experiment on development of robot gesture vocabularies, an approach to improve robot gestures in a guided way was envisioned. As it was already mentioned, interactive genetic algorithms have been applied in previous work by others for the goal of evolution based on aesthetic selection. Given that the improvement of gestures is in essence an aesthetic selection problem, interactive genetic algorithms were chosen to be used in the experiment.

Each gesture is encoded as a chromosome, consisting of genes. The chromosome represents the trajectory of the gesture, and each gene represents one point in space, also known as keyframe, that is a part of the trajectory. The evolution is achieved through application of genetic operators.

The algorithm used two genetic operators: single-point crossover and adaptive mutation. Selection algorithm for the crossover was queen selection algorithm developed by Stern et al. (2006).

Queen selection algorithm differs from the standard selection algorithm, in the way that there is a set of queens, consisting of N_Q members with highest fitness values. For mating purposes, one mate partner is selected from the queen set, while the other member is selected from the whole population. Mutation was performed on the level of keyframes, instead of on the level of single joint angles. That is, if a particular keyframe from a sequence is to be mutated, first a set of nearby keyframes was obtained and a mutated keyframe was selected from this set. This represents a new point in the gesture trajectory, thus replacing the old point.

Following parameters were selected empirically:

- Population size of $N = 26$ with six members chosen for presentation in each generation.
- Elitism was used, through which the member of the population with highest fitness is kept in the population.
- Maximum number of generations was set to 20.
- Number of queens in the queen set of $N_Q = 5$.
- Crossover probability of $p_c = 0.7$.
- Adaptive mutation was employed, meaning that the current probability of a mutation of a member was inversely related to its fitness value, that is $p_{m,i} = 1 - f_i$.

Initial population for an action consisted of pre-designed gestures $\mathbb{B}_{\mathbb{P}}(0) = \{A_{P,h}(0), h = 1, \ldots, N_P\}$, obtained through the experiment on robot gesture vocabularies, random modifications of the pre-designed gestures $\mathbb{B}_{\mathbb{M}}(0) = \{A_{M,h}(0), h = 1, \ldots, N_M\}$, and randomly generated gestures $\mathbb{B}_{\mathbb{R}}(0) = \{A_{R,h}(0), h = 1, \ldots, N_R\}$, satisfying relation $N_P + N_M + N_R = N$.

Experimental Setup

A graphical interface for the evolution experiment was developed, which implements the interactive genetic algorithm as a back-end for improvement of gestures, while an animated 3D model of the Aldebaran Nao robot was used to perform the evolved gestures. A screenshot of the program can be seen in Figure 4.13.

The first experiment was conducted with 26 students of the Department of the Industrial Engineering and Management, Ben-Gurion University of the Negev, Israel. Of those, 23 filled out the statistical survey (7 female, 16 male, mean age = 25.26, $\sigma = 1.84$). The participants were asked to rate the gestures for the following three meanings: "Hi, my name is Robowa" (Action 1), "Can I take away this cup?" (Action 2), and "Would you like me to clean the table?" (Action 3). Sizes of pre-designed gestures sets were $N_{P,A1} = 3$, $N_{P,A2} = 2$, and $N_{P,A3} = 1$ for Actions 1, 2, and 3, respectively. Sizes of randomly modified gestures sets were $N_{M,A_i} = 13 - N_{P,A_i}$.

During the experiment, a participant would select an action, after which six randomly chosen gestures from the whole population would began playing in parallel in simulated model. They were asked to assign fitness values to the presented gestures, based on how well they fit to the chosen action. After rating, they were instructed to click on the "Next" button and repeat the procedure for all following generations, until a pop-up message would instruct them that the evolution was over. The stopping criteria were either one of the success criteria, or that the generation number reached the allowed number of generations. The latter criterion was to prevent high familiarization with the gestures, which might hinder valid rating of gestures.

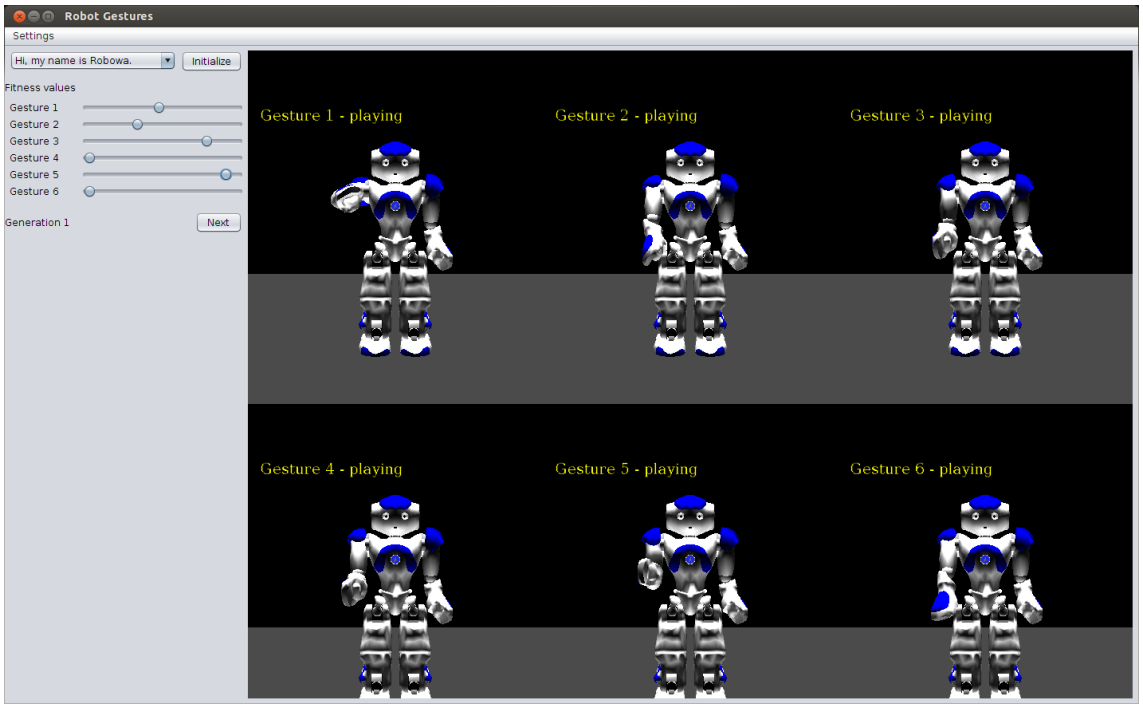


Figure 4.13: Experimental environment for evolution of robot gestures

Table 4.20: Results of the first experiment on gesture evolution with first success criterion

	Action 1	Action 2	Action 3	Overall
Valid runs	24	19	18	61
Successful runs	13	10	11	34
Success rate	54.17%	52.63%	61.11%	55.74%

Results

Three variations of a success criterion were defined to evaluate whether a particular run was successful: (1) to have a mean fitness of the generation above a particular threshold, in this case above 0.85, (2) to have at least one member of the population in the final generation having fitness value of 1, or (3) to have at least one member in any of the generations having fitness value of 1. Table 4.20 presents a summary of the experimental results.

The number of successful runs for the third action increased significantly between the second and the third variations of the success criterion. This is a result of the fact that some of the gestures were very long and therefore it was hard to associate them with the related actions. This was also reported by some of the experiment participants. The underlying cause was one of the pre-designed gestures, which was a longer sequence of keyframes. While the gesture itself was intended to be easily associable with its action, its length affected the duration of its descendants and this could cause the average length of gestures in the whole population to increase. Due to the nature of the algorithm, it is less probable to evolve a meaningful gesture, when its defined as a longer sequence of keyframes, compared to shorter sequences. Hence, in some experiment runs, the mean fitness value would increase up to a certain generation, followed by a decrease, once the duration of the gestures in the population increased (e.g. see Figure 4.14) To correct this problem,

Table 4.21: Results of the first experiment on gesture evolution with second success criterion

	Action 1	Action 2	Action 3	Overall
Valid runs	24	19	18	61
Successful runs	14	12	11	37
Success rate	58.33%	63.16%	61.11%	60.66%

Table 4.22: Results of the first experiment on gesture evolution with third success criterion

	Action 1	Action 2	Action 3	Overall
Valid runs	24	19	18	61
Successful runs	15	12	16	43
Success rate	62.5%	63.16%	76.19%	67.19%

that pre-designed gesture was removed from the initial population in the follow-up experiment.

Two examples of trends of mean fitness values of successful and unsuccessful evolutions are presented in Figures 4.15 and 4.16.

To verify the results obtained during the first experiment that the evolution was indeed successful and not a result of a participants' familiarization with the gestures, a second experiment was conducted with 7 participants from the Humboldt-Universität zu Berlin.

The experiment was modified, so that after the evolution was finished, due to either of the criteria, another set of six gestures were displayed to the participant. This set consisted of three best-ranked gestures from the first and from the last generation. After the gestures were rated, the ratio between the ratings of the gestures from the last and from the first generations was calculated. The ratio can be interpreted in the following way. If the ratio is equal to 1, then there is no difference in perception of the best gestures from the initial generation and from the last generation. The value of the ratio above or below 1 indicates that the gestures from the last generation were rated as better or worse compared to the best gestures from the first generation.

The verification was performed on the same actions as during the previous experiment: "Hi, my name is Robowa" (Action 1), "Can I take away this cup?" (Action 2), and "Would you like me to clean the table?" (Action 3). Following table 4.23 presents ratios for all three actions, as well as overall ratio.

The results show that the gestures evolved through the presented procedure are indeed perceived as better fitting to the actions, even when they are directly compared to the best rated gestures from the first generation. This shows that the ratings of the gestures do not artificially inflate due to getting accustomed to the experiment, but are getting better suited to the related actions. Another visible result is that the initial ratings and those obtained in the verification step are consistent.

Table 4.23: Results of the verification experiment on gesture evolution (standard deviation values reported in brackets).

	Action 1	Action 2	Action 3	Overall
Mean ratings, first generation, initial	0.38 (0.32)	0.41 (0.21)	0.41 (0.22)	0.4 (0.26)
Mean ratings, first generation, verification	0.33 (0.36)	0.55 (0.34)	0.44 (0.3)	0.42 (0.33)
Mean ratings, last generation, initial	0.96 (0.05)	0.92 (0.07)	0.96 (0.06)	0.95 (0.06)
Mean ratings, last generation, verification	0.88 (0.19)	0.76 (0.27)	0.92 (0.14)	0.87 (0.2)
Ratio	3.94 (4.22)	1.56 (0.74)	2.21 (0.57)	2.7 (2.68)

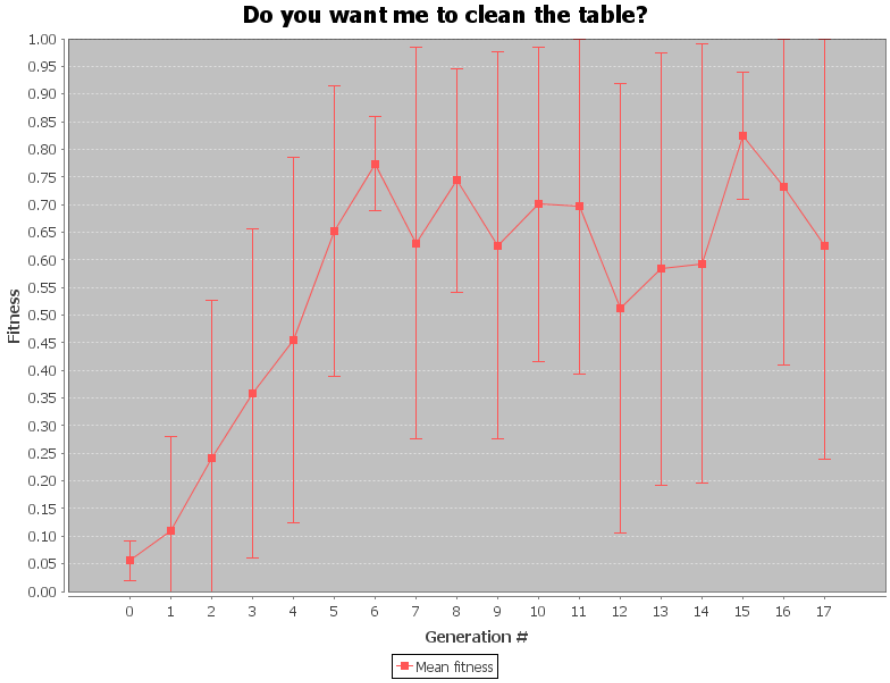


Figure 4.14: A decrease in mean fitness value for Action 7 during one experimental run.

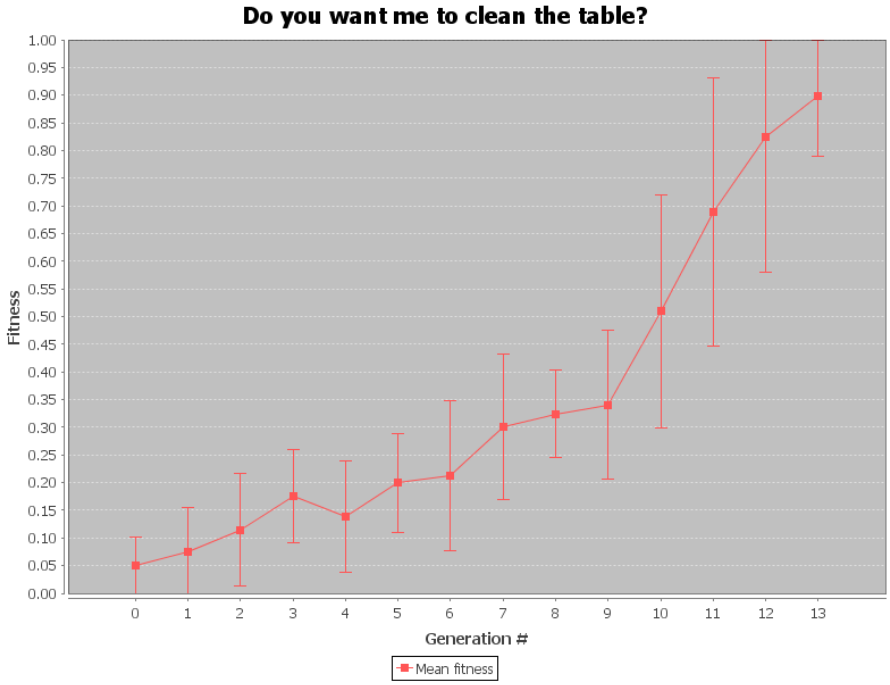


Figure 4.15: An example of the trend of mean fitness value for a successful evolution.

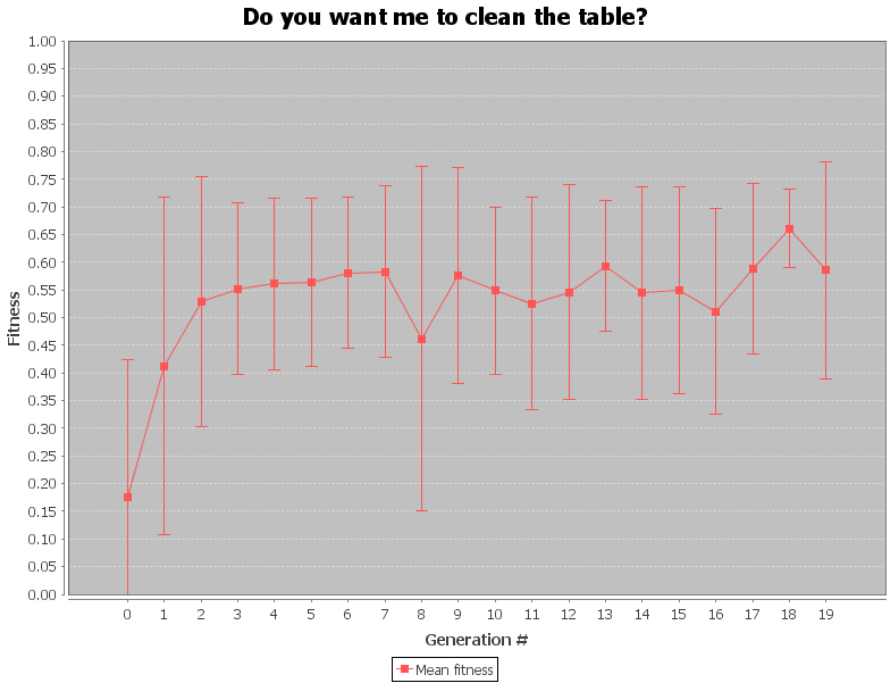


Figure 4.16: An example of the trend of mean fitness value for an unsuccessful evolution.

4.6 Discussion and Conclusions

4.6.1 Development of Gesture Vocabularies

Section 4.4 defined a methodology to select a set of gestures for a humanoid robot and a human, known as gesture vocabularies. The gestures were selected through two user surveys, in which the participants were asked to rank different gesture alternatives based on the measures of overall impression, speed and precision. Furthermore, they were asked to also rank the gestures from the best to the worst. Proposed methodology can be applied with different robot morphologies and different actions.

Lack of statistically significant results in few cases can lead to two explanations. One explanation is that the designed gestures are not well-known for the particular action, leading to the result where no particular gesture rated higher than the other. Another is that the gestures were not clearly readable on this particular robot. Proposed solution is to use interactive genetic algorithms to evolve a set of better gestures for these, as well as other actions, as it will be explained later.

To summarize, the first step is to define a set of actions for a particular task. Then, an initial set of robot gestures is defined for each task, with the idea of being similar to those performed by humans. These gestures are performed on a chosen robot and recorded. The next step is a survey where each participant is asked to rank these gestures from the best to the worst, and assign how precise the gestures are, is the speed of the gestures good and what their overall impression of the gestures is. Statistical analysis points out which gestures rated better than others. In case there is some improvement needed, a follow-up gesture evolution experiment can be run to evolve a set of more fit algorithms.

4.6.2 Evolution of Robot Gestures

The results of the experiment on robot gesture vocabulary showed the possible need to improve on the pre-designed gestures. As a way to achieve this, a method was envisioned based on aesthetic selection to improve the appearance of gestures performed by a humanoid robot. Interactive genetic algorithms were applied, based on their intended use in problems where it is hard, if not impossible to define a fitness function, such as in the problems of aesthetic selection.

The experimental results show that interactive genetic algorithms can be employed to improve robot gestures for predefined actions. The procedure starts with pre-designing gesture alternatives for one action, using them as a seed in the initial population, and then asking experiment participants to rate these gestures according to how well they fit to the current action. A follow-up validation step, in which three best rated gestures from the first and the last generations were presented in parallel in randomized order, confirmed that participants rated evolved gestures as better than those from the first generation, which confirms the good applicability of interactive genetic algorithms in evolution of gestures.

Chapter 5

Gesture Recognition and Disambiguation

5.1 Overview

This chapter relates to the topic of gesture recognition and disambiguation. In relation to the Figure 1.1, it covers the area of recognition of dynamic gestures and their disambiguation, as outlined in the Figure 5.1. Furthermore, the topic of learning of relations between observed pointing gestures and associated motor actions is explored through two experiments.

Sections 5.2 and 5.3 present a summary of gesture recognition methods and gesture representations, and reviews an application analysis for gestures in HRI and HCI. A short summary of the issues in gesture recognition is presented.

Two main areas are covered. Firstly, Section 5.4 introduces the gesture recognition and disambiguation framework. The main aim is recognition of dynamic arm gestures, while static gestures could be recognized to a certain extent, due to their different nature. The pre-processing pipeline, recognition method used and the disambiguation procedure are described in detail. The goal of this framework is to make the mechanisms of gesture recognition itself as abstract as possible, thus lowering the effect of interpersonal variations, in order to be able to use a simple recognition algorithm. Features that are ignored during the recognition can be used in the disambiguation step, in case they are relevant for this particular message. Section 5.5 presents the framework validation, including the required future validation. Performance of the recognition algorithm was measured using an extended testing method, which includes negative samples to simulate real-world environment. The results were obtained using a local gesture dataset, as well as the dataset compiled by Celebi et al. (2013).

Section 5.6 presents an example human-robot interaction scenario. In this scenario, a person can control the robot through gestures, giving it commands for what it should do.

Secondly, learning of deictic gestures is explored in Section 5.7, through use of internal models. They were employed in two experiments. In the first experiment, self-exploration lead to an execution of rotational movements, as a result of observed deictic gestures. The following experiment was based on learning by demonstration, where the initial execution mechanism was extended to translational movements.

The work presented in this chapter is partially based on following publications: Bodiroža et al. (2013a), Bodiroža et al. (2013b), Bodiroža and Hafner (2014), Doisy et al. (2013) and Jevtić et al. (2014).

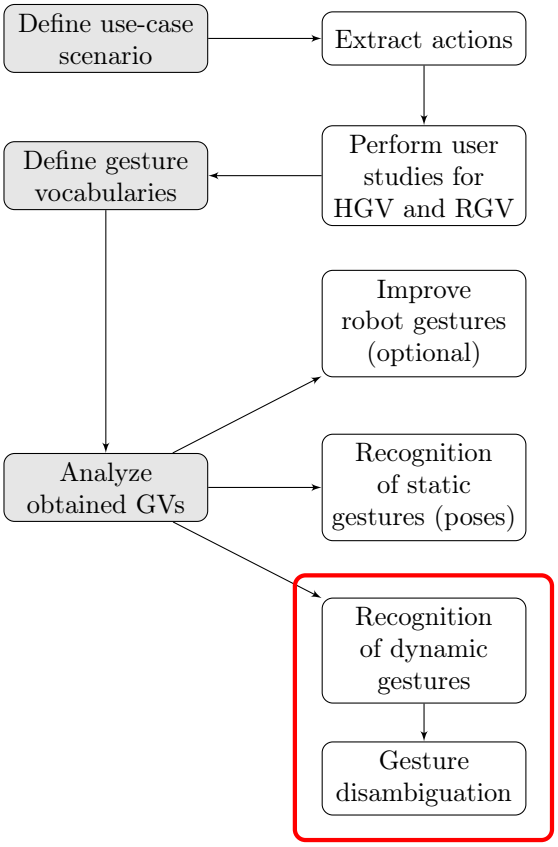


Figure 5.1: Research topics covered in Chapter 5

5.2 Background

Gestures are a prominently used modality in interpersonal interaction, usually accompanying speech. Therefore, they can be used as an intuitive interface for human-robot interaction, as well as human-machine interaction in general. In Chapter 3, certain aspects of robot's behavior have been identified in order to increase intuitiveness of interaction, such as enhanced feedback from the robot. Deictic gestures were employed by the robot to manipulate partner's attention. Gestures can also be used by the interacting partner in order to manipulate robot's attention and to indicate their own intentions and goals.

To enable use of gestures by humans, they need to be recognized by the robot. Gesture recognition is a process where data representing motion and pose information, which can be acquired through sensors, such as RGB and depth cameras, motion trackers, accelerometers and so on, is classified into one of the trained gesture classes. In order to have an effortless interaction, it is important to have robust gesture recognition, accommodating for inter- and intra-personal variations in gesturing. One of the prerequisites for robust gesture recognition are developed gesture vocabularies, which have been covered in Chapter 4.

Automatic gesture recognition, used in human-robot and human-computer interaction, has been an active area of research for at least three decades. As an example, Bolt (1980) developed a method for controlling an interface using speech and pointing gestures. However, there are still recurring issues that hinder its application in unconstrained, real-life scenarios. Most of the present algorithms for gesture recognition work well under certain assumptions, such as the location of the person in the scene or the direction of the gesture movement. Therefore, gesture recognition performs well while the adopted assumptions are satisfied, but fails when they are violated. The goal of this work is to reduce the number of required assumptions, in order to have both reliable and robust gesture recognition process.

Interpersonal variations in gesturing present a prominent issue, which lower the successfulness of gesture recognition. Furthermore, these variations are in some cases irrelevant with regard to the message to be conveyed. Inter- and intra-personal variations in gesturing is a recurring issue in gesture recognition. In previous chapter possible complexity of gesture vocabularies was made visible and the issues in their design and it presented which gestures people associate with particular actions (Bodiroža et al., 2012; Stern et al., 2006). Current approaches usually impose a limitation on the way how a gesture is performed, e.g., by enforcing a particular location in the gesture space where a gesture can be performed, or the direction of the gesture (e.g. see example gestures in work by Schlömer et al. (2008)). In some cases, these variations are not relevant for the exact associated meaning of the gesture. For example, a person might want to make a "circle" gesture, where it is not relevant whether the circle is large or small, or if it is gestured clockwise or counter-clockwise. Further issue is that people are not always standing still while they gesture in real-life conditions. However, these abstracted features are highly relevant in some cases. For example, if a user performs a "right to left swipe" gesture, the direction of the motion is important, but it is possible that it can be discarded during the recognition step. This can be addressed with the post-recognition disambiguation.

5.3 Gesture Recognition and Representations, Application and Advantages of the Proposed Framework

Approaches in gesture recognition were reviewed by Mitra and Acharya (2007), and more recently by Rautaray and Agrawal (2012). Different methods have been developed, e.g., hidden Markov models, dynamic time warping, finite state machines and time-delayed neural networks were employed for recognition of dynamic gestures. Gestures representations can also vary, such as absolute

or relative, 2D or 3D trajectories, or first derivative of the trajectories common in recognition with dynamic time warping (Corradini, 2001; ten Holt et al., 2007; Bodiroža et al., 2013a), quantizing the trajectory for recognition with a HMM (Yamato et al., 1992), using orientation histograms (Freeman and Roth, 1995), using a fixed-length histogram of features for detection using a SVM (Dardas et al., 2010) and so on.

As previously mentioned, the main idea behind the approach presented here is that gesture recognition in human-robot interaction should be robust, meaning that it should not be affected by particular environmental factors. These factors include the relative position and orientation of the person and the robot, physical size of the person, familiarity of the person with the gestures and the ease of their retraining.

There are prior efforts to make the gesture recognition more robust, such as the one presented by Sadeghipour and Kopp (2009), where the authors used internal simulations for gesture recognition and production. The work presented in this chapter introduces a concept of a comprehensive framework that aims to alleviate the problems typically encountered in the field of gesture recognition. The main aim is to recognize dynamic arm gestures, where static gestures could be recognized to a certain extent, due to their differing properties.

A recent work by Suarez and Murphy (2012) presented a review of state-of-the-art methods in hand gesture recognition and identified some of the still-present issues in gesture recognition. It is important to mention here that one of the identified issues is the location and orientation of a person, for example when they are not actively facing the sensor. Another issue is unfamiliarity of the person with a gesture set used by the system. This reiterates the need for increased tolerance of the gesture recognition algorithm, with regards to both inter- and intra-personal variances.

Wachs et al. (2011) analyzed possible applications of hand gestures in human-computer interfaces. Some of the identified costs and benefits of using hand gestures are price, responsiveness, user adaptability and feedback, learnability and low mental load, accuracy (detection, tracking and recognition), interaction space. The implemented framework addresses some of the mentioned issues. It is *responsive*, running in real time, due to the relatively simple algorithms which are used. The framework provides partial *user adaptability*. Because of the preprocessing, as explained in the Section 5.4.2, the recognition algorithm can be kept simple and it can work reliably with a low number of training samples. Furthermore, the preprocessing of the gesture trajectories reduces *mental load* and improves *learnability*, as some variations in performances of the same gesture are tolerated. *Interaction space* is not hindered, and the user can freely move around while gesturing.

Therefore, the developed approach alleviates to some extent the issues identified by Suarez and Murphy (2012) and takes into account the costs and benefits laid out by Wachs et al. (2011). Additionally, due to the low number of training gestures required, a particular gesture could be easily retrained if needed.

Noticeable differences can be observed between different persons and within one person, when they are performing the same gesture. Some of the features influencing the variations are the size of the gesture, its location in the gesture space, the speed of the hand(s) that perform the gesture and direction of the hand(s). These features can carry meaningful information, but in some cases they are irrelevant. For example, a person might want to indicate that an object had a circular shape and use a mimetic gesture that represents a circle. There are different possible ways to perform this gesture, which can be observed – it can be performed with the hand moving clockwise or counter-clockwise, or the resulting circle can vary in size between different persons. In other cases, these variations are relevant, and should be used to disambiguate between particular gestures, e.g. whether the round object was big or small, which could be indicated with the size of the "circle" gesture.

Table 5.1 summarizes some of the issues, identified in gesture recognition and relevant for the scope of this work, and proposed solutions presented in the rest of this chapter.

Table 5.1: Issues in gesture recognition

Issues	**Proposed solutions**
Interpersonal orientation	Frame of reference transformation
Gesture size	Normalization, standardization
Gesture location	Trajectory alignment (e.g., by alignment of first points or centroids)
Varying distribution of trajectory samples	Uniform resampling
Varying starting and ending points of a gesture	Bag of points

5.4 GRaD: A Framework for Gesture Recognition and Disambiguation

The goal of the framework for gesture recognition and disambiguation (*GRaD*) is to handle the issues that are common in gesture recognition in real-life scenarios.

Two key points are:

- gestures, regardless of whether the mentioned features are relevant or irrelevant, can be recognized using low number of training samples, and
- disambiguation, performed after recognition, discerns different gesture classes based on the training, such as "large circle" or "small circle" in case where two groups of training samples are present, one related to the former, and one to the latter class, or just a "circle" if there was only one class for a general "circle" gesture.

A low number of training samples is a highly important feature of the framework, enabling gesture recognition to become an easy to use, ubiquitous interface, having in mind that in adaptive systems, new gestures need to be added and current ones need to be modified while the system is in use.

Depth sensors have recently became affordable, alleviating the issues in gesture data acquisition imposed by computer vision. However, the data they provide is usually noisy and contains outliers. In addition, some features are not relevant for the recognition itself, while being relevant for disambiguation in some cases. In order to have robust gesture recognition, raw data is preprocessed to smooth the recorded trajectory and to remove particular features, such as gesture size, direction, velocity and velocity profile, location in the gesture space and starting and ending points of the gesture, that are not relevant for the recognition of the basic gesture. The idea is to make the gesture recognition invariant with regard to these features. Furthermore the recognition is performed on a trajectory which is standardized and aligned with the comparison gesture.

Figure 5.2 represents the general outline of the framework. A gesture trajectory is obtained through a sensor. The trajectory is preprocessed before the recognition. The output of the recognition algorithm is the identifier of the recognized gesture. In case there is some ambiguity, i.e., if this identifier is mapped to more than one gesture, then the identifier together with additional features is provided to the disambiguation algorithm to determine what is the observed gesture.

5.4.1 Data Acquisition

A gesture is represented as a time sequence $\mathbf{A}$ of 3D joint positions A_i, for $i = 1 \dots n$. This can be extended to include more than one joint, that is a gesture being represented as a set of time

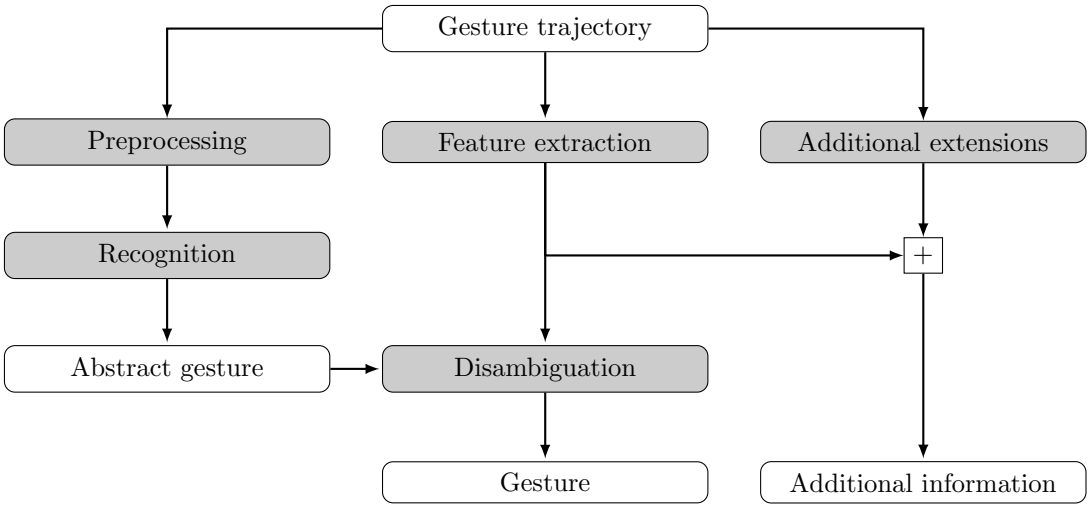


Figure 5.2: Proposed Gesture Recognition and Disambiguation (*GRaD*) framework

sequences $\mathbb{A} = \{\mathbf{A_{joint\ 1}}, \ldots, \mathbf{A_{joint\ m}}\}$, where m is the number of joints used for recognition. Joint positions can be obtained using different sensors, e.g. using RGB cameras, motion trackers or accelerometers. In this work, depth information is used to perform gesture recognition. Recent release of affordable sensors such as Microsoft Kinect and Asus Xtion PRO Live has allowed a wider acceptance of the depth-image technology by the scientific community. Human skeleton tracking algorithms (Shotton et al., 2011) use depth information to segment the human body into joints, which can be applied to gesture recognition. They provide 3D positions of body joints with a $30Hz$ frame rate.

5.4.2 Data Preprocessing

Data preprocessing consists of five stages and includes the following steps: frame of reference transformation, normalization or standardization, trajectory alignment, uniform resampling and direction-invariance. Throughout the rest of the work, the term "abstract gesture" will represent a gesture trajectory with discarded features. For example, the "swipe left" gesture, with the hand moving from right to left in front of the body, and the analogous "swipe right" gesture would be recognized as the same abstract gesture, that is they would be represented with a very similar direction-invariant trajectory.

Frame of Reference Transformation

Some gesture recognition algorithms tend to fail when the interacting partners are not standing still during gesturing. The reason is that the data provided by a depth sensor is grounded in the sensor's frame of reference. To enable the algorithm to compare observed gestures independently from the sensor relative position to the person, the coordinates of the joints are transformed from the sensor's frame of reference to a frame of reference fixed on the person. This frame of reference is computed for each skeleton data and centered on joint, that is selected as the anchor point.

Therefore, the first step is to ground the points of the observed trajectory, $\mathbf{A_i}$ to an anchor point fixed on a person, $\mathbf{A_{anchor}}$, e.g. to the point between the hip joints. Doing this, trajectories are observed in the person's frame of reference. However, this is still sensitive to the rotation of the person, relative to the axis drawn through the person's and the sensor's frames of reference.

To resolve this issue, the absolute orientation of the anchor point is obtained and the difference vector obtained in the first step is multiplied by the rotation matrix, **R**, providing the person's position- and rotation-invariant gesture representation (Bodiroža et al., 2013a). Similar approach was presented by Chaaraoui et al. (2014).

$$\mathbf{A_{i,invariant}} = (\mathbf{A_i} - \mathbf{A_{anchor}})\mathbf{R} \tag{5.1}$$

Trajectory Scaling

To enable gesture recognition with low number of training samples, all observed and trained gestures are scaled. The arm length of the person is used as a scaling factor. A vector representing the transformed position of a particular joint is divided with the arm length, resulting in the size of the gesture being scaled relative to the size of person who is performing the gesture.

Algorithm 1 Data scaling

Require: An array of n-dimensional points A.
Require: Arm length l

```
1: function ScaleTrajectory(A,l)
2:     for all Dimension d in A do
3:         for all Point a in A do
4:             a_d ← a_d/l
5:     return A
```

Trajectory Alignment

In order to make the recognition invariant with regard to the location where the gesture is performed within the gesture space, the observed sequence **A** needs to be aligned with the trained sequence **B**, to which the gesture is compared.

This alignment can be done in various ways. One way is to align the two sequences **A** and **B** with regard to their starting points A_1 and B_1. The displacement of the first point of the observed gesture A_1 from the first point of a known gesture B_1 is calculated as the difference vector, and this displacement is subtracted from all points of the sequence **A**, as presented in Algorithm 2, (Bodiroža et al., 2013a).

Algorithm 2 Trajectory alignment using the difference between first points

Require: A is an observed and B is a trained sequence.

```
1: function DiffAlignment(A, B)
2:     d_{x,y,z} ← b_{1x,y,z} − a_{1x,y,z}
3:     for all Point a in A do
4:         a_{x,y,z} ← a_{x,y,z} − d_{x,y,z}
5:     return A
```

Another approach is to find the centroid of sequence **B**, with $B_{CoG} = (\sum_{i=1}^{m} B_i)/N$, where m is the length of sequence **B** and subtract it from sequence **A**, as presented in Algorithm 3.

Testing showed that the recognition was in general better when using the centroid, especially taking in consideration when it is combined with the bag-of-points method, explained in Section 5.4.2, to make the recognition direction-invariant.

Algorithm 3 Trajectory alignment using the centroid

Require: A is an observed and B is a trained sequence.

1: **function** CentroidAlignment(A, B)
2: $sum_x, sum_y, sum_z \leftarrow 0$
3: **for all** Point b **in** B **do**
4: $sum_{x,y,z} \leftarrow sum_{x,y,z} + b_{x,y,z}$
5: $CoG_{x,y,z} \leftarrow sum_{x,y,z}/len(B)$
6: **for all** Point a **in** A **do**
7: $a_{x,y,z} \leftarrow a_{x,y,z} - CoG_{x,y,z}$
8: **return** A

Uniform Resampling

The next step is resampling of the trajectory to obtain one with a uniformly-distanced points. This is performed in order to diminish the effect of velocity profiles of different persons, as well as the difference in distribution of points when a periodic gesture (e.g. the "circle" gesture) has different starting points (e.g. beginning the "circle" gesture in the upper or the lower section). If this is not performed, due to these variations the observed and trained gestures could not always be well aligned and matched. The trajectory is resampled, so that the Euclidean distance between each two points is equal.

Direction Invariance

Final preprocessing step is performed in order to enable recognition of a specific case of gestures, which can be performed in different directions. For example, the "circle" gesture can be performed clockwise or counter-clockwise, and the "hand wave" could be performed starting from left to right, or from right to left.

A bag-of-points approach is used for making the recognition direction-invariant. Both the observed gesture and the one to which it is compared are represented using a bag-of-points model. A bag-of-points is a concept analogous to a bag-of-words in natural language processing. A gesture trajectory is stored in a multiset, an unordered sequence of points, where one point in space is allowed to have multiple instances. Each point from the observed gesture is matched to every point in the trained gesture in order to find the closest point. Distance between two points is determined using Euclidean distance as a measure. A "maximum mapped points" parameter defines what is the maximum number of points from one gesture that can be mapped to one point of another gesture. This parameter prevents the case of most of the points from one gesture being mapped into only one or two points of the other gesture. Figure 5.3 illustrates the described procedure.

5.4.3 Gesture Recognition

As presented in the literature review and in the introduction of this chapter, there are multiple approaches to gesture recognition. Depending on requirements, which are mostly defined by the intended use-case scenario, certain machine learning and pattern recognition methods can be applied.

Static and dynamic gesture recognition are two complementary areas. Static gesture recognition is concerned with the recognition of static postures, such as body or hand postures, while dynamic gesture recognition uses the dynamic component of a gesture for classification. For some areas,

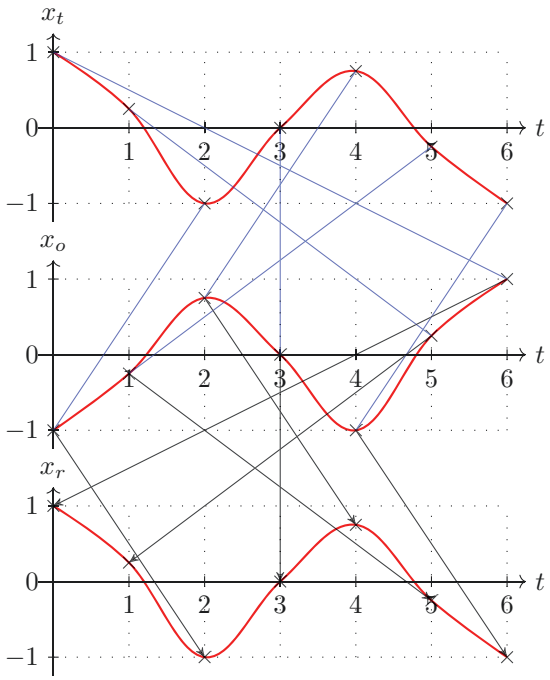


Figure 5.3: Illustration of direction invariance: remapping of points of the observed sequence x_o (second plot) by comparing them to the trained sequence x_t (first plot) in order to obtain the reconstructed sequence x_r (third plot) that best matches the trained one. The plots present only one dimension in time for illustration purposes. The observed sequence is the trained sequence performed in the other direction.

such as sign language recognition, both static and dynamic components of a gesture might need to be included.

The potential of use of the algorithm in real life situations is highlighted through its use on low number of training samples. Prior to the development, main consideration was that the approach should be practical, that is it should rely on a low number of training samples, while having high precision (i.e., low number of false positives) and high recall (i.e., low number of false negatives). On the other hand, it should be robust, so that it could recognize gestures of different persons, without needing to train it with each person separately.

These initial considerations influenced the development. The developed recognition algorithm relies on dynamic time warping, which performs nonlinear alignment of two sequences. This algorithm was initially developed for speech recognition by Sakoe and Chiba (1978). However, it was also successfully used in other areas, including gesture recognition, as presented by Corradini (2001).

Acquired data were partially preprocessed, as described in the previous section. Frame of reference transformation, normalization and trajectory alignment were applied, resulting in recognition which was location-, position-, scale- and speed-invariant.

Intuitive interaction with the robot requires generating a human gesture vocabulary. Improved robustness in real-life scenarios was required, having in mind possible integration with person-following, mobile robot (Doisy et al., 2013).

Dynamic Time Warping

The gesture recognition algorithm employs the multi-dimensional dynamic time warping algorithm in order to perform non-linear alignment of two time sequences, such as in ten Holt et al. (2007), Corradini (2001).

Let the sequence **A** be the observed gesture and the sequence **B** the known gesture.

Let $\mathbf{A} = \{A_1, A_2, \ldots, A_n\}$ and $\mathbf{B} = \{B_1, B_2, \ldots, B_m\}$ be two time sequences of lengths n and m, where $A_i = (A_{ix}, A_{iy}, A_{iz})$ and $B_i = (B_{ix}, B_{iy}, B_{iz})$. The DTW algorithm creates a dissimilarity matrix **D** of size $(n+1) \times (m+1)$. An element of the matrix $D_{i,j}, (i = 1 \ldots n, j = 1 \ldots m)$ stores the cumulative dissimilarity between the sequences **A** and **B** from 0-th elements of the sequences to the $(i-1)$-th element of sequence **A** and $(j-1)$-th element of sequence **B**, along a particular path through the matrix. The dissimilarity $d(A_i, B_j)$ between two aligned points in sequences is calculated as Euclidean distance:

$$\begin{aligned} d(A_i, B_j) &= \textsc{Euclid}(A_i, B_j) \\ &= \sqrt{(A_{ix} - B_{jx})^2 + (A_{iy} - B_{jy})^2 + (A_{iz} - B_{jz})^2} \end{aligned} \tag{5.2}$$

Additionally, two slope matrices **Si** and **Sj** of the same size as the matrix **D**, are maintained to track the allowable individual moves to reach a particular point (i, j) from initial point $(0, 0)$ in the dissimilarity matrix. If the maximum slope value is reached, the particular movement in the matrix only along the corresponding axis is prohibited, as explained below.

Dissimilarity matrix is initialized such that $D_{i,j} = +\infty$ for $i = 0 \ldots n, j = 0 \ldots m$, except when $i = j = 0$, and $D_{0,0} = 0$. Slope matrices are initialized to zeros. Afterward, it calculates the values of the elements $D_{i,j}$, for $i = 1 \ldots n, j = 1 \ldots m$. For each element $D_{i,j}$, it takes in consideration three elements $D_{i,j-1}$, $D_{i-1,j}$ and $D_{i-1,j-1}$ for potential predecessors. If the element at the position $(i, j-1)$ has lower value than the other two, and the value of $\mathbf{Si}_{i,j-1}$ is lower than the maximum slope value, then it is the predecessor, and the cumulative dissimilarity is:

$$\mathbf{D}_{i,j} = \mathbf{D}_{i,j-1} + d\left(A_{i-1}, B_{j-1}\right) \tag{5.3}$$

The slope matrix **Si** is updated such that $\mathbf{Si}_{i,j} = \mathbf{Si}_{i,j-1} + 1$, while $\mathbf{Sj}_{i,j} = 0$. In essence, the elements B_{j-2} and B_{j-1} correspond to the element A_{i-1}.

Else, the element $\mathbf{D}_{i-1,j}$ is taken as a potential predecessor and its value is compared to the other two. If its value is lower and the value of $\mathbf{Sj}_{i-1,j}$ is lower than the maximum slope value, then this element is selected as the predecessor, and the cumulative dissimilarity is:

$$\mathbf{D}_{i,j} = \mathbf{D}_{i-1,j} + d\left(A_{i-1}, B_{j-1}\right) \tag{5.4}$$

The slope matrix **Sj** is updated such that $\mathbf{Sj}_{i,j} = \mathbf{Sj}_{i-1,j} + 1$, while $\mathbf{Si}_{i,j} = 0$. In essence, the elements A_{i-2} and A_{i-1} correspond to the element B_{j-1}.

Finally, if neither is true, the cumulative dissimilarity is:

$$\mathbf{D}_{i,j} = \mathbf{D}_{i-1,j-1} + d\left(A_{i-1}, B_{j-1}\right) \tag{5.5}$$

Elements at position (i, j) of both slope matrices are set to zero. In essence, the element A_{i-1} corresponds to the element B_{j-1}.

As a result the elements of the last column of matrix **D** contain the dissimilarity values between the whole sequence **B** and every subsequence of **A**. A subset of these elements is matched to a threshold value to determine if they are within a range for the two sequences to be classified as similar. The subset represents the dissimilarity values between the whole sequence **B** and subsequences of **A**. The length of the subsequences is from manually set minimum length to the whole length of the **A**.

Algorithm 4 Dynamic time warping and minimum dissimilarity between warped trajectories

Require: An array of n-dimensional points $\mathbf{A} = A_1, \ldots, A_n$, observed gesture
Require: An array of n-dimensional points $\mathbf{B} = B_1, \ldots, B_m$, trained gesture
Ensure: Dissimilarity matrix $\mathbf{D}$ of size $(m+1) \times (n+1)$
Ensure: Minimum dissimilarity d

1: **function** DTW($\mathbf{A}, \mathbf{B}$)
2: $\mathbf{Si} \leftarrow [0]_{(m+1)\times(n+1)}$
3: $\mathbf{Sj} \leftarrow [0]_{(m+1)\times(n+1)}$
4: $s_{max,i} \leftarrow \text{smax}_i$
5: $s_{max,j} \leftarrow \text{smax}_j$
6: $\mathbf{D} \leftarrow \begin{bmatrix} 0 & \infty & \cdots & \infty \\ \infty & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \infty & 0 & \cdots & 0 \end{bmatrix}_{(m+1)\times(n+1)}$
7: **for all** i **in** 1, ..., (m+1) **do**
8: **for all** j **in** 1, ..., (n+1) **do**
9: **if** $\mathbf{D}_{i,(j-1)} < \mathbf{D}_{(i-1),(j-1)}$ & $\mathbf{D}_{i,(j-1)} < \mathbf{D}_{(i-1),j}$ & $\mathbf{Si}_{i,(j-1)} < s_{max,i}$ **then**
10: $\mathbf{D}_{i,j} \leftarrow \text{EUCLID}(B_{(i-1)}, A_{(j-1)}) + \mathbf{D}_{i,(j-1)}$
11: $\mathbf{Si}_{i,j} \leftarrow \mathbf{Sj}_{i,(j-1)} + 1$
12: $\mathbf{Sj}_{i,j} \leftarrow 0$
13: **else if** $\mathbf{D}_{(i-1),j} < \mathbf{D}_{(i-1),(j-1)}$ & $\mathbf{D}_{(i-1),j} < \mathbf{D}_{i,(j-1)}$ & $\mathbf{Sj}_{(i-1),j} < s_{max,j}$ **then**
14: $\mathbf{D}_{i,j} \leftarrow \text{EUCLID}(B_{(i-1)}, A_{(j-1)}) + \mathbf{D}_{(i-1),j}$
15: $\mathbf{Si}_{i,j} \leftarrow 0$
16: $\mathbf{Sj}_{i,j} \leftarrow \mathbf{Sj}_{(i-1),j} + 1$
17: **else**
18: $\mathbf{D}_{i,j} \leftarrow \text{EUCLID}(B_{(i-1)}, A_{(j-1)}) + \mathbf{D}_{(i-1),(j-1)}$
19: $\mathbf{Si}_{i,j} \leftarrow 0$
20: $\mathbf{Sj}_{i,j} \leftarrow 0$
21: curRow $\leftarrow 1$
22: curCol $\leftarrow 1$
23: **while** curRow $<$ m & curCol $<$ n **do**
24: **if** $\mathbf{D}_{curRow,(curCol+1)} < \mathbf{D}_{(curRow+1),(curCol+1)}$ & $\mathbf{D}_{curRow,(curCol+1)} < \mathbf{D}_{(curRow+1),curCol}$ & $\mathbf{Si}_{curRow,(curCol+1)} < s_{max,i}$ **then**
25: curRow $\leftarrow$ curRow+1
26: **else if** $\mathbf{D}_{(curRow+1),curCol} < \mathbf{D}_{(curRow+1),(curCol+1)}$ & $\mathbf{D}_{(curRow+1),curCol} < \mathbf{D}_{curRow,(curCol+1)}$ & $\mathbf{Sj}_{(curRow+1),curCol} < s_{max,j}$ **then**
27: curCol $\leftarrow$ curCol+1
28: **else**
29: curRow $\leftarrow$ curRow+1
30: curCol $\leftarrow$ curCol+1
31: $d \leftarrow \mathbf{D}_{curRow,curCol}$
32: **return** $\mathbf{D}, d$

The implemented algorithm is shown in Algorithm 4.

For the purpose of gesture recognition, the gesture trajectory is obtained through the skeleton tracker. It is stored as a sequence of (x, y, z) coordinates of the person's right hand. In the current implementation only right hand is tracked. However, the algorithm can easily track both hands. The height at which the Kinect is enables recording of a person at approximately shoulder level at $15fps$.

Gesture recognition is continuous with a sliding window of a certain length. Every time a new hand position is obtained, it is added to the end of a buffer. In case the buffer has reached maximum length, its first element is removed. This trajectory is then compared using DTW with learned trajectories.

5.4.4 Gesture Disambiguation

The work presented in this section is a theoretical concept of how a gesture disambiguation could be performed.

The main idea of the recognition algorithm was to strip particular features from gesture trajectories, therefore obtaining abstract gesture trajectories, in order to have a simple and robust gesture recognition. The goal was to enable easy training of the algorithm, relying on as low as one training sample for successful recognition, that would be robust and recognize the gestures that are not performed in the same manner and that might be performed by different people.

However, the disadvantage of the approach is that in some cases these features are relevant for correct classification of the observed gesture. This leads to induced ambiguity between these particular subsets of gestures, which are disambiguated with these features. A concrete example is recognition of "clockwise circle" and "counter-clockwise circle" gestures, or "swipe left" and "swipe right", whose only differentiating characteristic is the direction of the hand movement during gesturing.

As presented in Figure 5.2, outlining the overall *GRaD* framework, one part of the algorithm is feature extraction, which quantizes gesture features, such as size, velocity, direction and location. Following features could be used for disambiguation.

- Gesture size is represented with the difference of maxima and minima along the three dimensions of the trajectory.
- Gesture location in the gesture space is represented as the centroid of the gesture trajectory relative to an anchor point, e.g. the point between the hip joints.
- Beginning and end of the gesture trajectory are represented as those points relative to the anchor point.
- Gesture direction is represented as the vector along two most prominent dimensions, as the difference between the first occurring extremum and the second occurring extremum along a dimension.
- Simplified gesture velocity as a time duration of the particular gesture.

During the training of the recognition algorithm, dissimilarities of all 2-combination of abstract gestures classes are evaluated. In case the dissimilarity between two gesture classes is low, these classes are marked as requiring disambiguation. Otherwise, if the difference between a particular abstract gesture class and every other trajectory is above the threshold, then the output of the recognition algorithm is sufficient.

Linear classifiers could be employed to evaluate which features are potentially discriminating between the marked classes, and therefore could be used for disambiguation between them. In case

a recognizer produces as an output one of these two gestures, the disambiguation step is used to determine which one of the two gestures is true. Furthermore, this step could be extended in case there are multiple groups of similar gestures, e.g., if the training samples would contain both small and large "clockwise circle" and "counter-clockwise circle" as four separate gesture classes.

5.5 Performance Evaluation

The aim of the framework is gesture recognition robust to the intra- and inter-personal variations of gestures, such as size, location and speed of a gesture and the orientation and position of the person relative to the sensor or to the robot.

Furthermore, due to the preprocessing and the abstraction it provides to the gesture recognition algorithm, a dataset, containing gesture performances varying in size, speed, location and direction within the same class both with the person being static and moving around, should be specifically designed in order to fully test the effects of the proposed framework.

The evaluation was performed on the subset of the of the *GRaD* framework. However, gesture disambiguation requires future validation.

5.5.1 Evaluation of the Algorithm

The algorithm used OpenNI library for Microsoft Kinect for data acquisition, and it was implemented in C++ using Qt framework on Linux. The evaluation of the algorithm was performed on a local HU dataset, recorded and compiled by the author. In addition to this, the algorithm was tested using the dataset provided by (Arici et al., 2014).

HU Gesture Dataset

The local dataset was recorded and compiled by the author on 8 participants. It consists of following gestures: "Counter-clockwise circle", "Clockwise circle", "Swipe from right to left", "Swipe from left to right", "Stop", "Push", "Clean" and "Call". "Counter-clockwise circle" and "clockwise circle" represent a circular gesture in front of the body along the frontal plane in counter-clockwise and clockwise direction, respectively. "Swipe from left to right" and "Swipe from right to left" describe a motion where the right hand is moving from left to right, or from right to left, respectively, along the transverse plane. "Stop" gesture is performed by raising the right hand in front of the performer to shoulder height, holding it for a brief moment, and then lowering it back to the resting position. "Push" gesture is performed as if the performer was pushing a button in front of them at shoulder height with the right hand. "Clean" gesture can be seen as a gesture to remove something from the table and it is performed by repetitive motion of the right hand from and toward the body along the transverse plane. "Call" gesture is performed by raising the right hand laterally from a resting position in front of the body up next to the right shoulder, as if the performer is calling someone to join them. All gestures were performed with the right hand.

All subjects received a spoken description of a gesture they are about to perform, together with one exemplary gesture performance. All subjects performed 5 repetitions for each of 8 gesture classes. Furthermore, every participant provided 10 negative samples were included in the dataset, to approximate real-life conditions, where people might make unintentional hand movements, according to Pavlovic et al. (1997) taxonomy. These samples included both static postures and dynamic motions. In this case, the participants were instructed to make motions that regularly occur, but that do not resemble the recorded gestures, such as scratching the shoulder or the head, pointing to a side, or holding hands still. In some cases, 6 instead of 5 positive samples were recorded per participant per gesture class and these extra samples were left in the dataset. In total, there were 327 positive testing samples and 83 negative testing samples.

Table 5.2: Confusion matrix of the recognition of the gestures from the HU dataset without preprocessing, using dissimilarity threshold of $t = 75$ (rows – actual class, column – recognized class).

	CCW Circle	CW Circle	Swipe L	Swipe R	Stop	Push	Clean	Call	Neg
CCW Circle	**28**	0	0	0	0	0	0	0	**13**
CW Circle	0	**3**	0	0	0	0	0	0	**38**
Swipe L	0	0	**4**	0	0	0	0	0	**36**
Swipe R	0	0	0	**26**	0	0	0	0	**15**
Stop	0	0	0	0	**31**	0	0	0	**10**
Push	0	0	0	0	0	**18**	0	0	**23**
Clean	0	0	0	0	0	0	**5**	0	**36**
Call	0	0	0	0	0	0	0	**17**	**24**
Negative	0	0	0	0	0	0	0	**6**	**77**

The sensor was placed at the height of $h = 1.4m$ and the person performing the gestures for the training dataset was standing at the distance of $d = 2.5m$ from the sensor. Two datasets were recorded. The first dataset contains recordings from 8 participants, which were standing still facing the sensor while gesturing. The second dataset contains recordings from 2 participants and it is specific in a sense that the participants were moving backward and forward while performing the gestures.

Training was performed using 2 training samples per gesture class. Every testing sample was compared against both training samples of every gesture class using dynamic time warping. As a result of the comparison, the higher dissimilarity value was returned.

The gesture trajectories were saved in raw format, without any preprocessing of the data, storing the absolute (x, y, z) joint positions and the rotation matrices of the joints. Each recording is 120 frames long, to account for the preparation and retraction phases during the recording.

Evaluation on the HU Gesture Dataset

The dissimilarity threshold value was empirically set to $t = 0.15$ when the gesture was preprocessed using scaling. This states that the average distance between aligned points in training and testing samples must not be larger than 7.5% of the maximum size of the gesture along x, y and z axes in order for testing gesture to be classified to the class of the training gesture sample. In the case where gesture is not scaled, the threshold value was empirically set to $t = 75$. Positions of only right hand was used for comparison.

Following confusion matrices, where rows are actual gesture classes and columns are recognized gesture classes, represent the results with different settings of the preprocessing pipeline:

- Table 5.2: No preprocessing.
- Table 5.3: Frame of reference transformation.
- Table 5.4: FoR transformation, gestural alignment.
- Table 5.5: FoR transformation, gestural scaling.
- Table 5.6: FoR transformation, gestural alignment and scaling.

Important statistics are number of true positives (i.e., correctly classified positive gestures), true negatives (i.e., non-classified negative gestures), false positives (i.e., misclassified positive or negative gestures) and false negatives (i.e., non-classified positive gestures). using the notion of

Table 5.3: Confusion matrix of the recognition of the gestures from the HU dataset with frame of reference transformation, using dissimilarity threshold of $t = 75$ (rows – actual class, column – recognized class).

	CCW Circle	CW Circle	Swipe L	Swipe R	Stop	Push	Clean	Call	Neg
CCW Circle	**26**	0	0	0	0	0	0	0	**15**
CW Circle	0	**26**	0	0	0	0	0	0	**15**
Swipe L	0	0	**12**	0	0	0	0	0	**28**
Swipe R	0	0	0	**17**	0	0	0	0	**24**
Stop	0	0	0	0	**28**	0	0	0	**13**
Push	0	0	0	0	0	**35**	0	0	**6**
Clean	0	0	0	0	0	0	**22**	0	**19**
Call	0	0	0	0	0	0	0	**40**	**1**
Negative	0	0	0	0	0	0	0	**3**	**80**

Table 5.4: Confusion matrix of the recognition of the gestures from the HU dataset with frame of reference transformation and alignment of gestures, using maximum alignment difference of $d = 0.2$ and dissimilarity threshold of $t = 0.15$ (rows – actual class, column – recognized class).

	CCW Circle	CW Circle	Swipe L	Swipe R	Stop	Push	Clean	Call	Neg
CCW Circle	**33**	0	0	0	0	0	0	0	**8**
CW Circle	0	**36**	0	0	0	0	0	0	**5**
Swipe L	0	0	**35**	0	0	0	0	0	**5**
Swipe R	0	0	0	**36**	0	0	0	0	**5**
Stop	0	0	0	0	**40**	0	0	0	**1**
Push	0	0	0	0	0	**41**	0	0	0
Clean	0	0	0	0	0	0	**33**	0	**8**
Call	0	0	0	0	0	0	0	**38**	**3**
Negative	0	0	0	0	**4**	0	**34**	**3**	**42**

Table 5.5: Confusion matrix of the recognition of the gestures from the HU dataset with frame of reference transformation and scaling of gestures, using maximum alignment difference of $d = 40$ and dissimilarity threshold of $t = 75$ (rows – actual class, column – recognized class).

	CCW Circle	CW Circle	Swipe L	Swipe R	Stop	Push	Clean	Call	Neg
CCW Circle	**34**	0	0	0	0	0	0	0	**7**
CW Circle	0	**37**	0	0	0	0	0	0	**4**
Swipe L	0	0	**25**	0	0	0	0	0	**15**
Swipe R	0	0	0	**31**	0	0	0	0	**10**
Stop	0	0	0	0	**39**	0	0	0	**2**
Push	0	0	0	0	0	**41**	0	0	0
Clean	0	0	0	0	0	0	**35**	0	**6**
Call	0	0	0	0	0	0	0	**41**	0
Negative	0	0	0	0	0	0	0	**26**	**57**

Table 5.6: Confusion matrix of the recognition of the gestures from the HU dataset with frame of reference transformation, and alignment and scaling of gestures, using maximum alignment difference of $d = 0.2$ and dissimilarity threshold of $t = 0.15$ (rows – actual class, column – recognized class).

	CCW Circle	CW Circle	Swipe L	Swipe R	Stop	Push	Clean	Call	Neg
CCW Circle	**39**	0	0	0	0	0	0	0	**2**
CW Circle	0	**41**	0	0	0	0	0	0	0
Swipe L	0	0	**38**	0	0	0	0	0	**2**
Swipe R	0	0	0	**38**	0	0	0	0	**3**
Stop	0	0	0	0	**41**	0	0	0	0
Push	0	0	0	0	0	**41**	0	0	0
Clean	0	0	0	0	0	0	**35**	0	**6**
Call	0	0	0	0	0	0	0	**40**	**1**
Negative	0	**1**	0	0	**1**	**2**	**2**	**3**	**74**

Table 5.7: Summary of evaluation results

	TPR	TNR	FPR	FNR	PPV	NPV
Table 5.2	40.37	92.77	7.23	59.63	95.65	28.31
Table 5.3	63.00	96.39	3.61	37.00	98.50	39.80
Table 5.4	89.30	50.60	49.40	10.70	87.69	54.55
Table 5.5	86.54	68.67	31.33	13.46	91.59	56.44
Table 5.6	95.72	89.16	10.84	4.28	97.20	84.09

true positive rate $TPR = \frac{TP}{P} = \frac{TP}{TP+FN}$, true negative rate $TNR = \frac{TN}{N} = \frac{TN}{TN+FP}$, false positive rate, $FPR = \frac{FP}{N} = \frac{FP}{FP+TN}$, false negative rate $FNR = \frac{FN}{P} = \frac{FN}{TP+FN}$, precision (or positive predictive value) $PPV = \frac{TP}{TP+FP}$, and negative predictive value $PPV = \frac{TN}{TN+FN}$.

The results from tables 5.2-5.6 are summarized in Table 5.7.

Positive and negative predictive values provide information about the ratio of the actual true positives and negatives, compared to all classified positive and negative outcomes. It can be seen that all five settings of the preprocessing pipeline didn't affect the positive predictive value significantly. However, the effect of the complete preprocessing can be easily seen in the increasing negative predictive value (see summary of the results from Table 5.6), which means the algorithm is better at discarding negative samples when the preprocessing is used.

Sehir University Gesture Dataset

The algorithm was also evaluated on the dataset compiled by Celebi et al. (2013). The dataset consists of following, self-descriptive gestures: "both hands pull down", "both hands push up", "both hands zoom in", "both hands zoom out", "left hand pull down", "left hand push up", "left hand swipe right", "left hand wave", "right hand pull down", "right hand push up", "right hand swipe left" and "right hand wave". The authors originally evaluated their algorithm, called weighted dynamic time warping, on a subset of those gestures: "left hand push up", "left hand pull down", "left hand swipe right", "right hand push up", "right hand pull down" and "right hand swipe left", as summarized in the table 5.8. Results obtained from the recognition algorithm presented here are presented in Table 5.9. It can be seen that both approaches provide comparable result. However, negative samples were not present in the dataset. Positions of both right and left hand were using during the comparison.

Table 5.8: Confusion matrix of the recognition of the gestures from the Sehir University gesture dataset, data represented in percentage, results from a paper by Celebi et al. (2013) (rows – actual class, column – recognized class).

	R push U	L push U	R pull D	L pull D	R swipe L	L swipe R
R push up	**100**	0	0	0	0	0
L push up	0	**100**	0	0	0	0
R pull down	0	0	**100**	0	0	0
L pull down	0	0	0	**85**	**15**	0
R swipe L	0	0	0	0	**100**	0
L swipe R	0	0	0	0	**15**	**95**

Table 5.9: Confusion matrix, using the recognition algorithm presented in this chapter, using maximum alignment difference of $d = 0.6$ and recognition threshold of $t = 0.58$ (rows – actual class, column – recognized class).

	R push U	L push U	R pull D	L pull D	R swipe L	L swipe R
R push up	**100**	0	0	0	0	0
L push up	0	**97.31**	**2.68**	0	0	0
R pull down	0	0	**97.5**	**1.65**	0	**0.83**
L pull down	0	0	0	**100**	0	0
R swipe L	0	0	0	0	**100**	0
L swipe R	**2.59**	0	0	**1.72**	0	**95.69**

5.6 Gesture-based Control of a Person-following Robot

For a human-robot interaction to take place, a robot needs to perceive humans. The space where a robot can perceive humans is restrained by the limitations of robot's sensors. These restrictions can be circumvented by the use of external sensors, like in intelligent environments; otherwise humans have to ensure that they can be perceived. With the robot platform presented in this section as an example interaction, the roles are reversed and the robot autonomously ensures that the human is within the area perceived by the robot. This is achieved by a combination of hardware and algorithms capable of autonomously tracking the person, estimating their position and following them, while recognizing their gestures and navigating through environment[1].

5.6.1 Background

The space where human-robot interactions can take place is restricted by the limited perceptual capabilities of robot's sensors and/or the lack of mobility of the robot. By using external sensors and creating intelligent environments, e.g., as in the work of Morioka et al. (2004), robots are able to follow humans. Recent release of the affordable depth sensors lead to similar solutions that do not rely on external sensors, as presented by Doisy et al. (2012). Natural interaction techniques, such as gesture recognition, assume that the relative position of the sensor and the human remains unchanged during the interaction (ten Holt et al., 2007; Corradini, 2001). Such techniques need to be improved to be used in less constrained scenarios (Bodiroža et al., 2013a). This part presents a robot platform benefiting from the latest development in person following and gesture recognition to expand the space where HRI can take place.

[1] Aleksandar Jevtić and Guillaume Doisy were responsible for person tracking, person position estimation, person following and system integration. Author's contribution was in gesture recognition.

5.6.2 Framework for Extended Human-Robot Interaction

Person Tracking

Person tracking is performed using the Kinect sensor and the Microsoft Kinect SDK. The limited 57° horizontal viewing angle of the sensor does not allow for person tracking 360° around the robot. To overcome this limitation a visual servoing control law was developed to rotate the pan-tilt mechanism (PTM) and the Kinect sensor in the direction of the tracked person. Experimental results showed that this visual control law ensures continuous person tracking when the person is moving around the robot. The latency of the Kinect sensor is too high to allow for continuous tracking when both the robot and the person are moving: in the case of fast movements the system is not reactive enough and loss of tracking can occur. This problem was solved by adding a term in the PTM control law which compensates the angular speed of the robot given by the odometry (i.e. making the PTM rotate with a speed opposite to the angular speed of the robot). For a formal description of the control law, please refer to the work of Doisy et al. (2012).

Person Position Estimation

It is possible to compute the absolute position of the tracked person by using a series of frame of reference transformations, if the position of the person in the Kinect's frame of reference, the rotation angle of the PTM and the absolute position of the robot estimated from odometry are known. In order to eliminate the noise due to the vibrations of the robot structure a $15cm$ jitter filter and a low-pass filter based on a $1.5m/s$ speed threshold and a $1g$ acceleration threshold are applied to the position estimation. The accuracy of the position estimation method was evaluated by comparing the estimated position of a person with its known ground path during 10 trials. Results show that this method is able to estimate the position of a person with an average error of $4.2cm$ (s.d. $2.6cm$) and a maximum error of $10.4cm$.

Person Following

With the knowledge of the position of the person the robot can perform person following. Three algorithms were developed and implemented on the robot platform:

- direction-following: the robot always goes in the direction of the person it follows (Morioka et al., 2004).
- path-following: the robot reproduces the path of the person. With this technique the robot can follow a person in an environment with low-height obstacles, e.g. chairs or tables, assuming that the path taken by the person is free of obstacles (see Figure 5.5).
- adaptive following: with an a priori obtained map of the environment the robot continuously computes and takes the shortest path to reach the position of the followed person (Doisy et al., 2012).

Position-Invariant Gesture Recognition

Gesture recognition employs the multi-dimensional dynamic time warping algorithm, similar to the approach used by ten Holt et al. (2007) and Corradini (2001), which is used to align and compare two time sequences. In this work, coordinates of human joints, provided by the Kinect sensor, are used. In order for the algorithm to be independent from the relative speed, position and orientation of the robot and the person, the joints coordinates are transformed from the Kinect's frame of reference to the person's frame of reference. This frame of reference is defined by the

center point of the left and right hip joints and a unit vector orthogonal to the ground. In essence, the algorithm represents the subset of that presented previously in this chapter. For recognition results, see Tables 5.10-5.17.

Robot Platform robuLAB10

The robot platform is a Robosoft's robuLAB10, customized with a rigid structure, which supports a TRACLabs Biclops PTM (see Figure 5.4). The mechanism can support a maximum payload of $4kg$, which was enough to carry the Kinect sensor. The robot, the Kinect sensor and the PTM were controlled by a laptop powered by an Intel quad-core i7 Q740 CPU with 4 GB of RAM.

5.6.3 Person Following by a Robot Controlled with Gestures

To demonstrate capabilities of the described robotic platform, a person-following application was developed and tested, which was controlled by gestures. It used the specific capabilities of the platform: person tracking to keep the sensor oriented toward the person, person position estimation and path following to make the robot follow the person and dynamic gesture recognition to control the robot.

System Behavior

Three robot states were defined: "passive tracking", "active tracking" and "following". To switch between the states four command gestures were created: "start active tracking", "stop active tracking", "start following" and "stop following". In "passive tracking", the PTM and the robot were static. In "active tracking", the robot remained static but the PTM was moving to direct the sensor toward the person. In "following", the robot followed the person using the path-following algorithm.

Experiments

To test the system's performance, 10 trials were conducted. The user was instructed to signal to the robot to start to actively track them, then to follow them in an environment with obstacles and along a path similar to the one displayed on Figure 5.5 and finally to stop the following and return to the "passive tracking" state. Performance was measured using four metrics: 1) task completion, 2) number of correctly recognized commands (hit), 3) number of incorrectly recognized commands (false positive) and, 4) number of undetected commands.

Results

In all the trials, the user was able to initiate and to stop tracking and following using gestures with a goal of guiding the robot to the desired location. In 10 trials, out of 40 commands (4 commands per trial), only 5 were missed and had to be repeated and there were no false positives. These results demonstrate the feasibility of controlling the robot using gestures. In particular, it has been proven that it is possible to simultaneously track a person in motion, estimate their position and recognize gestures using a single depth sensor mounted on a mobile platform without the use of external sensors.

5.7 Internal Models for Gesture Recognition

Internal models represent a theoretical concept, consisting of a pair of inverse and forward models, represented in figure 5.6. Depending on a problem, an inverse model can predict a motor command

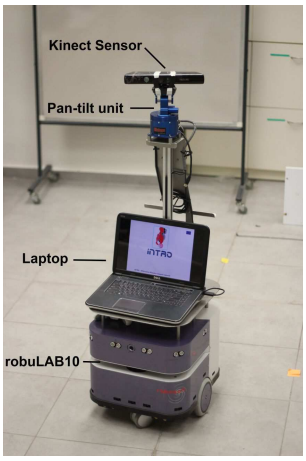


Figure 5.4: Robot platform robuLAB10.

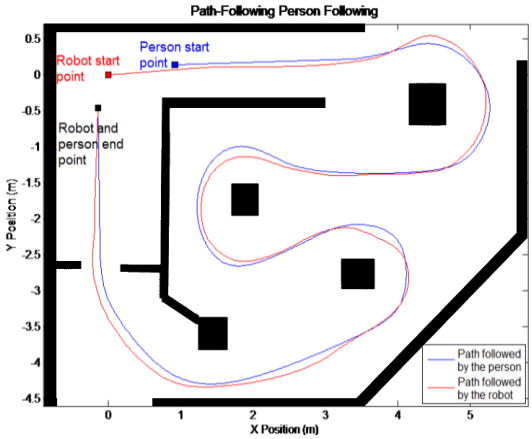


Figure 5.5: Example of a path-following algorithm.

M_t that leads the system from the current state S_t to a desired state S_{t+1}, or, based on the observed change of states from S_t to S_{t+1}, predicts an appropriate motor command M_t. The forward model performs an internal simulation and predicts the state S^*_{t+1} that would be the result of the motor command M_t in the state S_t.

5.7.1 Inverse and Forward Models for Action Execution

Two experiments are presented, which demonstrate the possibility of using internal models for learning the relation between motor actions and changes in sensory information, as consequences of the action. The first experiment uses self-exploration for learning the relation between rotation and perceived change in the position of a person's hand. The learned mapping can be used to train a robot how to rotate in direction indicated with a hand motion. The second experiment uses learning by demonstration to train a robot how to reach a position on the ground, indicated with a pointing gesture.

Similar to this approach, Dearden and Demiris (2005) perform learning of forward models for action execution. In their work, a mobile robot observes the motion of its gripper while sending to

it random motor commands. A forward model is obtained, that establishes the connection between motor commands and the changes in the visual space caused by those motor commands. This way, the system is able to imitate human movements.

Akgün et al. (2010) show how an action generation mechanism can be used for action recognition. They developed an online recognition system, that was able to recognize a reaching action before it was fully executed.

Haruno et al. (1999) and Wolpert and Kawato (1998) present evidence for development of multiple, tightly-coupled inverse and forward models. The forward model predicts the result of the motor command generated by the inverse model. The selection of the best inverse-forward model pair is done through comparison of predictions of all forward models to the expected result. Schillaci et al. (2012) used multiple internal models to perform recognition of human behavior, where each internal model encodes an action. Blakemore et al. (2000) present how a difference in the prediction of the forward model and the perceived sensory input helps a person discriminate a self-induced sensation (e.g. self-movement of the eye) from a sensation induced by others (e.g. moving the eye by pressing on the eyelid). Furthermore, Takemura and Inui (2011) present a model for the development of internal models for reaching movement, inspired by infant development.

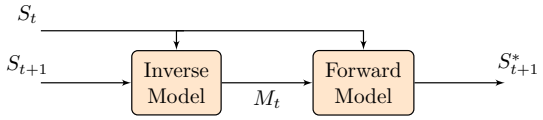


Figure 5.6: Internal models. An inverse model predicts a motor command M_t that leads the system from the current state S_t to S_{t+1}. A forward model predicts the state S^*_{t+1} based on the current state S_t and the motor command M_t.

Internal Models for Perception of Deictic Gestures

As previously mentioned in Section 2, gesture recognition is based on detection of the position (static) or the motion (dynamic) of human body parts, usually arms, hands and legs. It has numerous applications, including sign-language for hearing-impaired people, computer interfaces, natural and intuitive human-robot interaction, gaming industry, system remote control, among others. Various tools have been used for gesture recognition, based on the approaches ranging from statistical modeling, computer vision and pattern recognition, image processing, connectionist systems, etc. (Mitra and Acharya, 2007).

However, learning from self-exploration to observe and reproduce certain actions from gestures of others has not been done before, apart from work of Dearden and Demiris (2005). The proposed model learns to execute actions associated with directional gestures. The learned motor controls are a result of a motor babbling process, during which a random motor command is generated and a change in the sensory input is observed and stored.

5.7.2 Learning Motor Control for Rotation Based on Arm Movement

The proposed model uses as sensory situation the detected coordinates of the arm of a person as tracked by a Kinect. The motor commands are fixed movements of the mobile platform.

In this model, the goal is to perform the fusion of sensorimotor information. To achieve this, the system needs to collect, for each time step, a vector of the form:

$$(x, y, z); M \tag{5.6}$$

where (x, y, z) represent the coordinates of the hand detected by the Kinect and M represents a random motor movement. This movement can be either in the left-right plane, performed by the robot as a rotation for a random angle, within $[-27°, 27°]$ or in the up-down plane, performed as a tilt angle in the Kinect within $[0°, 16°]$ (ranges were selected to always have the person's upper body visible).

Once a database of these associations is collected, it can be used as either a forward or an inverse model, depending on the question asked.

The inverse model predicts a motor command M_t, when presented with a change in the sensory situation from S_t to S_{t+1}. The forward model, given the current sensory situation S_t and a motor command M_t predicts the new sensory situation after the execution of the command, S^*_{t+1}. In the proposed application, that is a gesture-controlled robot, during the learning process random motor commands induce changes in sensory situations, that is in the position of the hand. The model associates the performed motor command M_t with the sensory situation before the execution S_t and after the execution S_{t+1}. During execution the model performs a search for a motor command M_t that matches the sensory change, only this time induced by a person by moving their hand.

It could be said, that the model learns how to execute actions. It learns how the self-motion corresponds to changes in the world, which is then applied during the execution, when it is required to reproduce an action that resulted in the observed change. The resulting behavior can be also seen as an attention manipulation system, where the robot turns following the motion of a hand.

Experiments and Results

Two experiments were performed to learn the mapping between the motor commands and changes in the sensory situation. A robot platform robuLAB, displayed in Figure 5.4 and described in previously in Section 5.6.2 , was used in the experiments. In the first experiment, the association of "up" and "down" gestures and Kinect's up and down movements was formed, while in the second experiment the association of "left" and "right" gestures and the rotation of the robuLAB platform was learned. The following description shows the outline of both experiments, and "the platform" represents either the Kinect or the robot platform. In the former case, the motor commands were tilting the Kinect, while in the latter they were rotation of the robot platform around the z-axis.

A person stood in front of the platform at the beginning of the experiment. The Kinect was tracking the location of the person's right hand, provided by the Microsoft's Kinect SDK. During the learning stage, illustrated in Figure 5.7, the platform was performing motor babbling, which was a generation of random motor commands. Every motor command induced a change in the sensory situation, that is the change of the 3D location of the person's hand. This change, represented with the (x, y, z) vector of the hand movement in Kinect's frame of reference, corresponds to the rotation angle of the platform. In other words, if the platform rotates to the left, the hand will be seen as moving right. However, during the execution phase, illustrated in Figure 5.8, if the platform perceives the hand moving right, it should rotate to the right, instead of left. During the learning process, initial rotation of the platform and the generated random motor command, represented as a rotation angle, was stored, as well as the 3D position of the hand before and after the rotation.

A k-nearest neighbors search algorithm was used for the implementation of the inverse model. Theoretically, the initial state S_t represents the initial position of the hand, and the next state S_{t+1} the new position of the hand, resulting from the person's movement. The predicted motor command M_t represents the rotation angle of the robot that it needs to perform in order to compensate for the motion of the person's hand. However, in order to make the implementation more robust with regards to the person's location in space, S_t and S_{t+1} are combined and represented as the difference vector of the new and the initial location of the person's hand.

When a movement of the person's hand is observed, the perceived displacement of the hand is

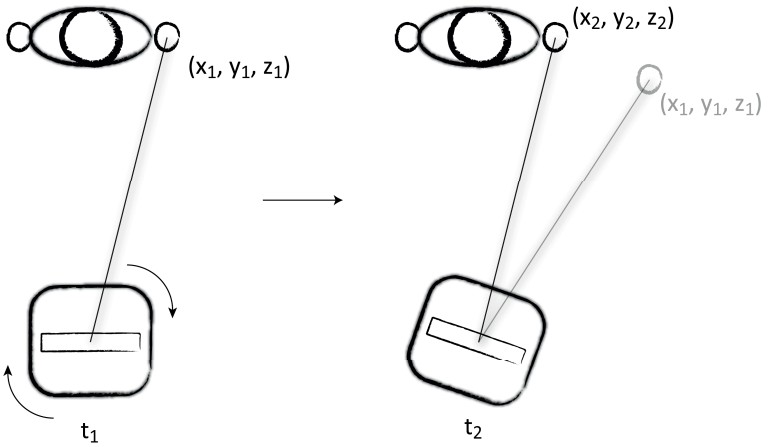


Figure 5.7: Representation of the learning process

used to predict a motor command that will compensate for the motion of the arm. As a result, the robot performs a rotation. This shows that inverse models can be used for perception of dynamic gestures, indicating the corresponding rotation, without prior coding of this behavior. The behavior of the robot can be indeed seen as recognition of a dynamic gesture and execution of the corresponding, learned action.

Testing of the algorithm was done on the robuLAB robot platform, with differential drive being used to perform rotations. Training was performed with 60 points, obtained as a result of motor babbling. Testing was performed with hand movements to the left or to the right for 25 times. On average, absolute hand displacement of the user was $(x, y, z) = (0.43, 0.24, 0.06)m, s.d. = (0.13, 0.12, 0.05)$ and the error of the prediction resulted in the absolute mismatch between the initial hand position and the hand position after the rotation of $(x, y, z) = (0.08, 0.24, 0.06)m, s.d. = (0.06, 0.12, 0.04)$. The results show that this approach can be used for learning motor control for rotation based on the depth data of the user's hand.

This part presented learning and execution of an inverse-forward model pair to execute actions associated with directional gestures. While only the inverse model was trained, the forward model could be easily added and used to predict the hand location after the robot's movement. This information can be used for error measurement of the prediction and refinement of the initial motion, if the error is higher then a certain threshold. An improvement of the proposed model, presented in the following section, was developed, as an extension toward understanding of deictic gestures, with the goal of learning a control strategy for the robot to guide itself to a specific location, indicated by a person pointing.

5.7.3 Learning Motor Control for Rotation and Translation Based on Pointing Gestures

This work presents an action execution system that uses as foundation basic sensorimotor schemes. These schemes are learned as a product of the interaction of an agent with its environment. The presented work is inspired by child development, and employs learning by scaffolding. In the experiments presented, a mobile agent learns an association between changes in its sensory perception and its movement, guided by the demonstrator. In previous section an agent was presented whose goal was to explore its environment and learn the relation between its rotation and the sensory

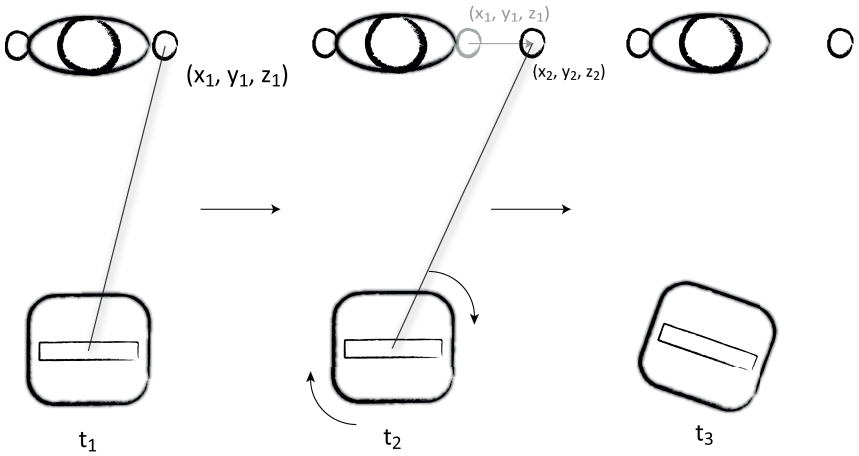


Figure 5.8: Representation of the execution process

perception caused by this motion. Once it acquired this knowledge, the agent was capable of performing a mirror action to match an observed gesture. This is seen as a first step toward learning motor control strategies for a robot control task.

As a follow up, the same mechanism is extended towards rotation and translation motions, that is movements in two dimensions on the ground plane, while including human guidance for learning appropriate motor commands. During training, the demonstrator points to a spot on the floor and then guides the robot to that location, obtaining sensorimotor pairs. After learning, the agent is capable of executing the necessary motor commands to go to a location to which the demonstrator is pointing. Contrary to analytical approaches in pointing gesture recognition, which usually find a line that goes through the arm and locates its intersection with the ground, the proposed approach employs learning with scaffolding. In this approach, the agent locates shoulder and hand joints using a RGB-D sensor.

During learning, the robot stores sensory input (x, y, z), representing the difference between the person's hand and the shoulder, after which it is guided by a person with a controller to the specified point, storing the motor output (θ, r), representing rotation and translation motor commands. Internal models learn an association between the stored gesture and the trajectory that followed. The agent uses a pair of interconnected self-organizing maps, to learn sensorimotor schema that map external stimuli, that is a particular configuration of the human arm, into a motor command that would bring the robot to the predicted point on the floor. The first self-organizing maps describes the topology of the sensory input, and the second of the motor output. Each node of one map is connected to every node of the other map. Modified Hebbian learning is employed to learn these connections, which are later used to find a motor command related to a particular position of the demonstrator's arm.

Results presented in Table 5.19 show that the robot can guide itself to positions within the trained area. In general, it is observed that the points that were closer to the person who was pointing had smaller errors, compared to those further away from the person, resulting in better recognition of those deictic gestures. This was not a surprising effect, considering that the same difference in pointing angle when pointing to a far point results in larger floor distance, compared to when pointing to a near point. A disadvantage of the proposed approach is that a person needs to have the same relative position to the robot as during the training. This should be addressed as a part of future.

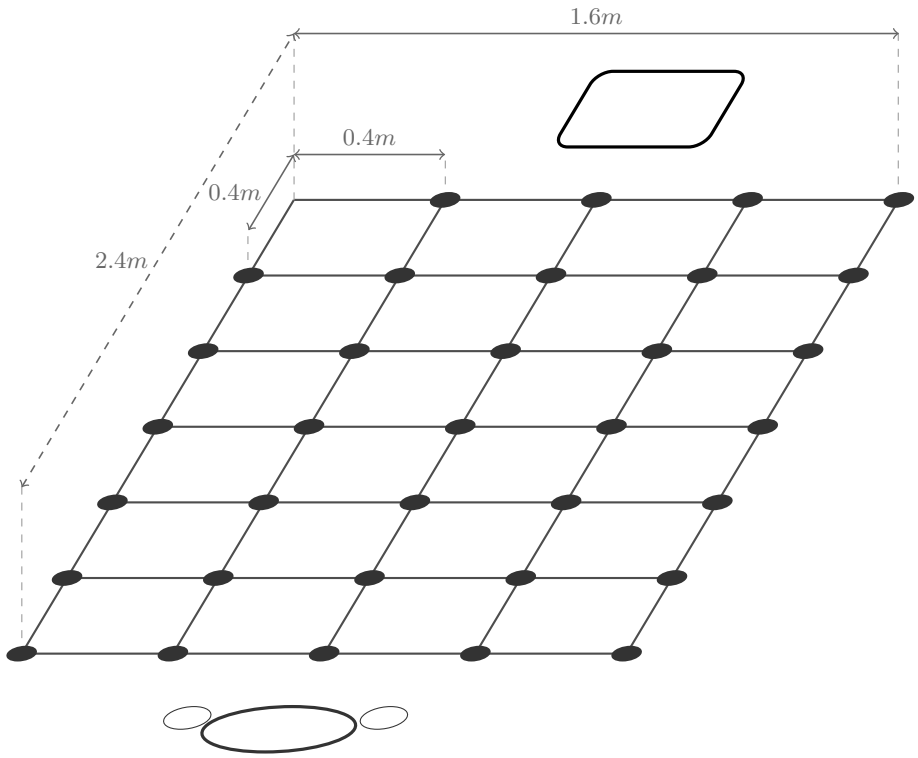


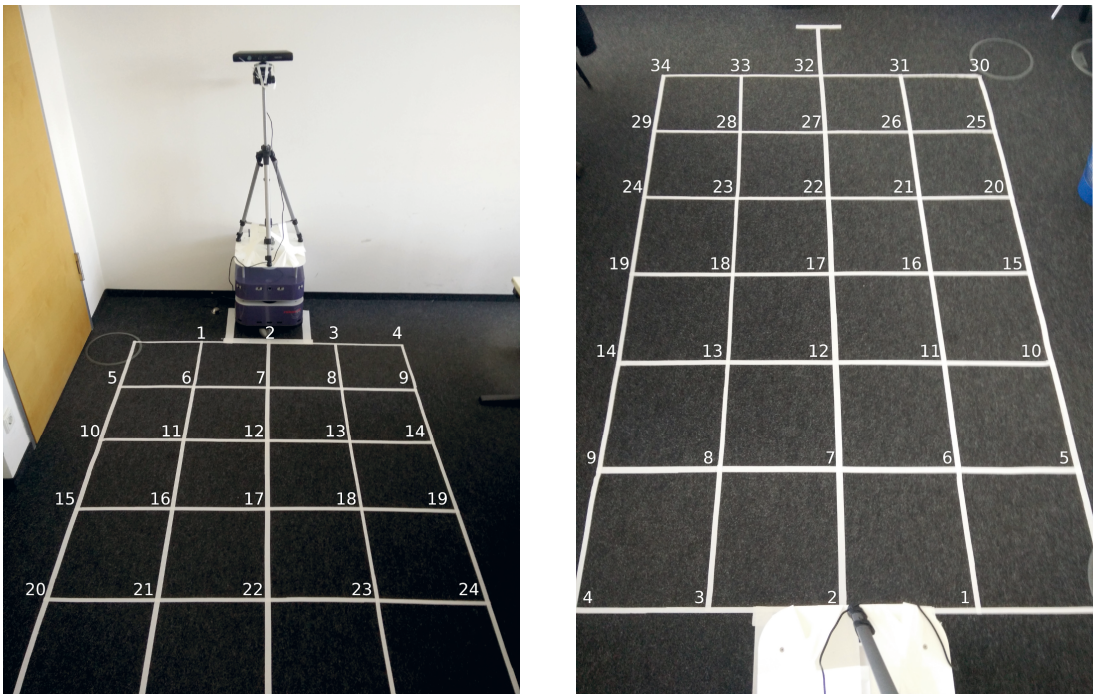
Figure 5.9: Illustration of the experimental setup for motor control learning. The location of the teacher is represented with an ellipse, while the position of the robot is marked with a rounded rectangle.

5.8 Discussion and Conclusions

Sections 5.4 and 5.5 presented a framework for recognition and disambiguation of trained gestures, and the results of the evaluation of the proposed approach. Recognition was based on dynamic time warping. Main issue was to develop a recognition system which uses one-shot learning, meaning that a very low number of training samples is used, while keeping the system robust to changes. This was possible due to a pre-processing pipeline, in which the gestures were recorded in user's frame of reference, as opposed to sensor's frame of reference. Furthermore, they were aligned to the training instances by their starting position, as well as normalized, in order to lower the effect of the gestural size on the recognition. Recognition results show that the system, which was trained on gestures performed by one person with two repetitions per gesture class, can recognize gestures performed by different persons, without having them instructed in detail how they should perform those gestures. Furthermore, a theoretical outline was presented for disambiguation of gestures, as well as particular features that could be used for the purposes of disambiguation.

Section 5.6 presented an example scenario, where gestures are used for control of a mobile robot. As a proof of concept, the recognition algorithm was tested under different environmental conditions, such as the person standing on different locations relative to the robot, or performing recognition while the robot was in motion.

Section 5.7 presented two experiments that were exploring the possibilities of learning relations between sensory changes and motor actions. In the first experiment, the robot learned how to



(a) Person's point of view (b) Robot's point of view

Figure 5.10: Experimental setup for motor control learning.

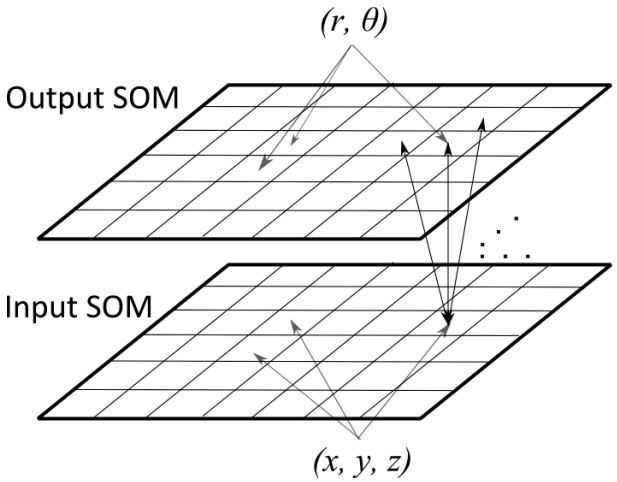
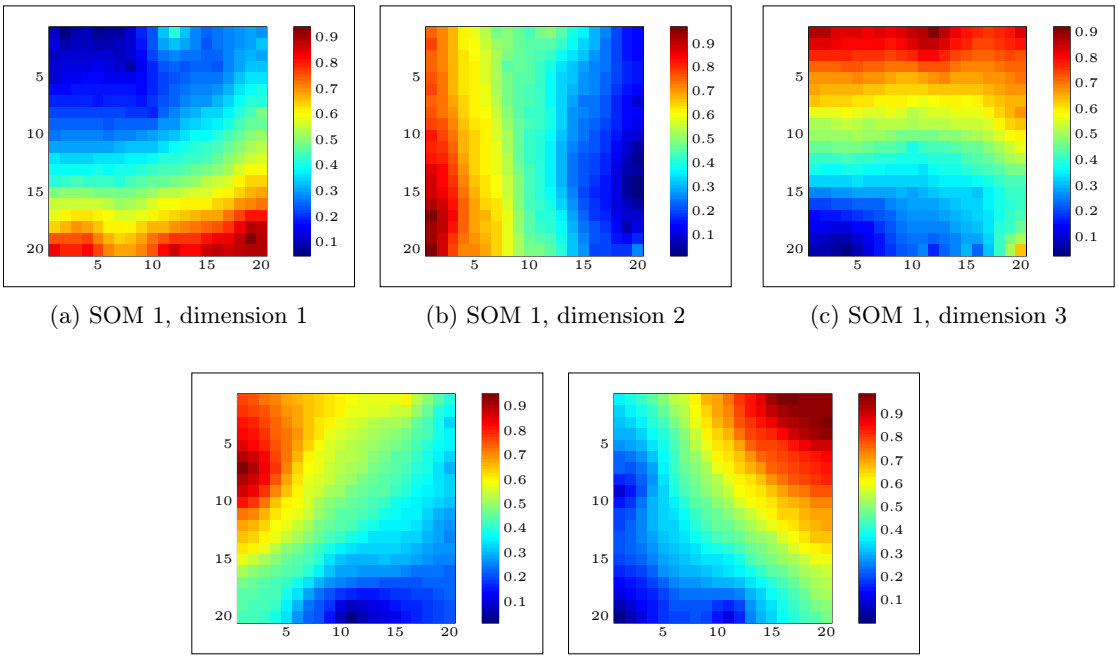


Figure 5.11: Illustration of the interconnected self-organizing maps.

rotate around its z-axis, as a response to a pointing gesture. The learning was done through self-exploration, during which the robot was tracking how does its own rotation affect the perceived location of the right hand of a person standing in front of it. In the second experiment, the robot would observe a person pointing to a particular location on the ground in front of the robot, followed by the person guiding the robot to the pointed location. Through establishing the relation



(a) SOM 1, dimension 1 (b) SOM 1, dimension 2 (c) SOM 1, dimension 3

(d) SOM 2, dimension 1 (e) SOM 2, dimension 2

Figure 5.12: Weights of input dimensions of self-organizing maps encoding perceived pointing (SOM 1) and related motor commands (SOM 2)

between the position of the arm and the motor commands, the robot was able to navigate to pointed locations on the ground without explicitly implementing this feature.

Table 5.10: Fixed PT-M with no displacement

No. ($1m$)	**Hit**	**Miss**	**CR**	**FA**
G1	0	10	30	0
G2	10	0	25	5
G3	5	5	30	0
G4	10	0	20	10
No. ($1.5m$)	**Hit**	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	10	0	29	1
G3	9	1	30	0
G4	10	0	30	0
No. ($2m$)	**Hit**	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	10	0	30	0
G3	9	1	30	0
G4	10	0	30	0
No. ($2.5m$)	**Hit**	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	10	0	30	0
G3	10	0	30	0
G4	10	0	30	0
No. ($3m$)	**Hit**	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	10	0	28	2
G3	8	2	30	0
G4	10	0	30	0
No. ($3.5m$)	**Hit**	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	10	0	22	8
G3	2	8	30	0
G4	10	0	30	0
No. ($4m$)	**Hit**	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	5	5	26	4
G3	2	8	30	0
G4	10	0	21	9

Table 5.11: Fixed PT-M with displacement of $0.5m$ to the left

No. ($2m, 3m, 3.5m$)	**Hit**	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	10	0	30	0
G3	10	0	30	0
G4	10	0	30	0
No. ($2.5m$)	**Hit**	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	10	0	27	3
G3	7	3	30	0
G4	10	0	30	0

Table 5.12: Fixed PT-M with displacement of $0.5m$ to the right

No. ($2m, 2.5m$,$3m, 3.5m$)	**Hit**	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	10	0	30	0
G3	10	0	30	0
G4	10	0	30	0

Table 5.13: Fixed PT-M with displacement of $1m$ to the left

No. ($2.5m$)	**Hit**	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	10	0	27	3
G3	7	3	30	0
G4	10	0	30	0
No. ($3m$)	**Hit**	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	10	0	26	3
G3	6	4	30	0
G4	10	0	30	0
No. ($3.5m$)	**Hit**	**Miss**	**CR**	**FA**
G1	4	6	30	0
G2	2	8	29	1
G3	3	6	30	0
G4	9	1	10	20

Table 5.14: Fixed PT-M with displacement of $1m$ to the right

No. ($2.5m$)	**Hit**	**Miss**	**CR**	**FA**
G1	1	9	30	0
G2	10	0	28	2
G3	10	0	30	0
G4	10	0	24	6
No. ($3m$)	**Hit**	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	10	0	26	3
G3	6	4	30	0
G4	10	0	30	0
No. ($3.5m$)	**Hit**	**Miss**	**CR**	**FA**
G1	8	2	30	0
G2	10	0	29	1
G3	9	1	30	0
G4	10	0	30	0

Table 5.15: Moving PT-M, distance of $2m$ from the robot

No. $(0°, -45°, -90°)$	**Hit**	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	10	0	28	2
G3	8	2	30	0
G4	10	0	30	0
No. $(45°, 90°, -135°)$	**Hit**	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	10	0	29	1
G3	9	1	30	0
G4	10	0	30	0
No. $(135°)$	**Hit**	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	10	0	30	0
G3	10	0	30	0
G4	10	0	30	0

Table 5.16: Robot moving forward toward the person

	Hit	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	10	0	29	1
G3	9	1	30	0
G4	10	0	30	0

Table 5.17: Robot moving perpendicular to the person

	Hit	**Miss**	**CR**	**FA**
G1	10	0	30	0
G2	8	2	25	5
G3	0	10	30	0
G4	10	0	23	7

Table 5.18: Robot person-following control

	Miss
Start tracking	2
Start following	0
Stop following	1
Stop tracking	2

Table 5.19: Average error (standard deviation in brackets) in *cm* during testing of the motor control learning (Field 1 in the first row was not included in testing due to room characteristics).

	1	**2**	**3**	**4**	**5**
1		0 (0)	0 (0)	3 (0)	5 (0)
2	8 (0)	5.8 (1.79)	2.6 (5.81)	2 (0)	9.2 (5.4)
3	3 (0)	2 (0)	0 (0)	3 (0)	3 (0)
4	3 (0)	5 (0)	7 (0)	4 (0)	7 (0)
5	1 (0)	13.6 (0.89)	6.8 (4.02)	15 (0)	7.5 (2.52)
6	0 (0)	14 (3.39)	1 (2.23)	9 (2.23)	4 (0)
7	12 (0)	7.2 (4.38)	4.8 (1.1)	20 (0)	23 (0)

Chapter 6

Conclusions and Future Work

The initial goals of the thesis were to address the questions of (a) methods for development of gesture vocabularies, (b) recognition and synthesis of gestures, and (c) approaches to gesture disambiguation. As an identified prerequisite, the question of attentional models was also tackled. Based on the research questions, two aims of the thesis were identified – (a) to establish an effective method for development of gesture vocabularies, and to (b) develop a method for reliable and robust gesture recognition, together with possibility of gesture disambiguation as a post-recognition step.

6.1 Attentional Models

Chapter 3 tackles the question of attentional mechanisms as a prerequisite for effective human-robot interaction. It presents a developed attentional model for a humanoid robot for exploration of robot's environment, employing a representation of short-term memory. This resulted in a seemingly natural exploration of the environment and attentiveness to located salient features. The particular implementation consisted of motion, face and object detection filters for detection of salient features, with a rule for assigning and decreasing saliency of each detected feature, in order to simulate inhibition, habituation and forgetting processes. These filters were used as a bottom-up attentional process, while four behaviors simulating top-down influence were implemented, which were using data provided by the ego-sphere to drive the interaction. An experiment with human participants was performed to understand how do different behaviors of the robot affect the way participants perceive it.

The main outcomes of the experiment was that use of attentional models provides a way to have a seemingly natural and intuitive interaction. Furthermore, the results identified the way different qualities of robot's behaviors influence the way others perceive these behaviors. Once there was a mechanism to direct robot's attention to interesting points in its surrounding, gestural interaction could be introduced.

6.2 Gesture Vocabularies

First main aim of the thesis was to introduce an effective method for development of gesture vocabularies. Chapter 4 presents the efforts toward this aim. The work on gestures in human-robot interaction consists of an initial analysis of the use-case scenario, development of human and robot gesture vocabularies for the particular scenario, and development of gesture recognition algorithms. Types of gestures present in the obtained human gesture vocabulary naturally affect the development of the gesture recognition algorithm. For example, one of the characteristics might

be which body parts are included during gesturing, or what kind of arm gestures are present (e.g., iconic or deictic). A museum guide robot might need to focus on pointing gestures when people are inquiring, while a dance tutor robot might require recognition of dynamic, whole-body gestures, as well as analysis of static dance poses.

This aim is addressed in chapter 4, which defined gestures and gesture vocabularies and presented an approach for development of human and robot gesture vocabularies. To verify the proposed approach, two experiments were conducted, in order to obtain both kinds of gesture vocabularies. Successful development of both gesture vocabularies displayed the effectiveness of the presented methods. This scenario is based on interaction between customers and waiters in a restaurant or a bar. Regarding robot gesture vocabulary, arm gestures were designed for a humanoid robot Nao based on gestures observed in human gesture vocabulary and real, interpersonal interactions in bars and restaurants. It was shown that most

The selected use-case scenario was a robot waiter scenario, based on interaction between a customer and a waiter occurring in a typical restaurant scenario. Actions were identified, based on observations of real-life interpersonal interaction between a customer and a waiter . Actions can be seen as customer-side, when they are initiated by a customer, and waiter-side, when they are initiated by a waiter.

Gestures for each of the customer-side actions were obtained through a user survey, where participants were presented with an action and asked to produce a gesture they associated with that action. Most of the identified actions identified as a part of human gesture vocabulary resulted in either one or two gestures, having a joint agreement level higher than 50%, meaning that more than half of the participant associated and gestured one or two gestures for one action.

Gestures for waiter-side actions were designed, based on gestures common in interpersonal interaction, and based on the gestures observed in the human gesture vocabulary, if an analogous customer-side action exists. These gestures were then presented in a user survey to participants, who were asked to rate the displayed gestures. Statistical analysis of the data was performed to see which gestures are consistently being highly rated, if such gestures existed. For majority of actions, there were gestures which were rated as fitting to their respective actions.

However, the results of the experiment on robot gesture vocabularies led to a question of whether robot gestures can be improved in cases where they were not rated high enough. A follow-up experiment was devised based on interactive genetic algorithms. The goal of the experiment was improve a robot gesture or a set of gestures for a chosen action through the use of evolutionary methods. In particular, an improved gesture would be that, which a participant or a group of participants would consistently rank the evolved gestures as better, compared to the initial gesture or gestures. An initial population of gestures, consisting of designed gestures, random modifications of those gestures and randomly generated gestures, was set and a random subset of those was presented to a human rater in parallel. The rater assigned fitness values to the presented gestures, according to rater's perception of how fitting the gestures were to the selected action. The fitness values served to assign probability to the gestures to be selected for reproduction, where offspring should ideally be better fitting gestures. The results of the experiment displayed that this iterative procedure can be well suited to produce improved gestures. This was shown with a post-evolutionary step, where the rater was asked to rate best rated gestures from the first and from the last generation.

6.3 Gesture Recognition and Internal Models

Second main aim of the thesis was that of development of a reliable and robust gesture recognition. Following the work on gesture vocabularies, chapter 5 introduced a framework for gesture recognition and disambiguation. This chapter consists of two main topics – a procedure for recognition of trained gestures together with a theoretical foundation for gesture disambiguation procedure,

and learning the relation of gestures and their meanings through internal models. A recognition algorithm was developed, based on dynamic time warping. This method performed non-linear alignment of two time sequences in order to reduce any time-induced differences in those sequences. The aligned sequences were compared using a distance function to measure their dissimilarity. A pre-processing pipeline was designed to further increase the robustness of the recognition, through elimination of some features of the gesture trajectories, resulting in what was called "abstract trajectories". The results showed satisfying recognition rate of selected gesture classes. An example scenario was presented in section 5.6, in which a person could control a robot through gestures.

Section 5.7 handles the second main topic of the chapter. This section presented two experiments how can the relation between a gesture and its associated meaning be learned. In particular, the first experiment presented an approach for issuing a rotate command to the robot by pointing left or right. In this case, the command was not explicitly programmed, but it was rather the learning approach that was programmed. Through this, the robot was able to learn how much it needs to rotate when issued a pointing gesture to the left or to the right. This was based on the prior self-exploration, where the robot was observing the displacement of a hand of the person standing in front of it while it was rotating around. The second experiment also employed pointing gestures, but in this case, they were used for navigating a robot to a particular location in front of the robot. Testing of the learning showed that it is possible to learn the relation between a particular arm position, representing a pointing to a particular location in the environment and the command that needs to be issued in order to move to that location.

6.4 Future Work

Due to the diverse topics covered in the thesis, this section will cover particular research directions in separate subsections.

6.4.1 Attentional Models

Interesting future directions would be to explore different approaches for dynamic weight assignment for different filters. It could be also useful to extend the system to include more filters on the Nao robot (e.g. for audio localization), as well as to port the approach to other robot platforms. It would be interesting to see how these attentional models would rate on other, non-humanoid platforms. Additionally, the presented *full interaction* behavior, consisting of *exploration*, *interaction* and *interaction avoidance*, can be applied to more complex scenarios, and it should be explored further. Providing giving visual and auditory feedback to the participant is of extreme importance for increasing the intuitiveness of the interaction and the user satisfaction, and this should be included in the future.

Another interesting research question could ask what is the proper movements speed a robot might exhibit in order to be perceived as not dangerous or with good reaction times.

6.4.2 Gesture Vocabularies

Overall, general remarks from the participants indicate further morphological requirements. The robot should be equipped with sound feedback, as well as arms which have rotatable wrist and movable fingers.

To test the human understanding of gestures in the robot gesture vocabulary, an experiment should be performed where participants would be asked to pair a selection of few gestures from this set with a selection of few actions. The expected outcome is that they would correctly match

the gestures with actions and this would prove that the chosen gestures can be clearly understood (e.g., a gesture for one action would not be recognized as a gesture for some other action).

Steels (1998) introduced the concept of language games – a method for evolution of a common language vocabulary between two agents. In more recent time, Steels and Spranger (2012) provides an overview of the additional games, such as action games and description games. A game consists of a population of individual agents, a context and a communicative purpose.

Similarly, in Schulz et al. (2006) an approach for evolution of spatial language is described. In this work two robots are exploring the locations and sharing between themselves the information of where these locations are and which words they are using as names of these locations. The names are randomly created from primitive syllables. The results show that two agents are capable of evolving a spatial language that can be learned and used for describing spatial information.

Following the same line of thought, an interesting approach to development of the gesture vocabularies are gesture games, in which a gesture vocabulary would be evolved between participants. This approach is highly relevant for a theoretical explanation of how could a common gesture vocabulary in a culture evolve.

In this case, a robot and a human teacher, or two robots, would watch each other performing random movements, which they use to describe a shared concept (i.e., an object of joint attention). These movements are randomly generated from movement primitives, such as raising a hand, followed by mutual agreement on the correspondence between a generated gesture and the shared concept.

6.4.3 Gesture Recognition and Internal Models

The approach to gesture recognition here is robust to variations in environment and persons. However, it does not tackle the question of context in which the gesture is performed, such as the current interaction context, or the referent object of the gesture.

Another focus should be on extending the framework to include the recognition of static gestures, as well as to include recognition of hand poses. Furthermore, as outlined in Figure 5.2, additional features could be explored, such as information about the gestural plane, as some of the 3D gestures are performed in a 2D plane, as indicated by Berman and Stern (2012).

Regarding the use of internal models for robot guidance, there is an issue of sensitivity to the relative position of a person towards a robot. This issue could be approached by inclusion of additional training features, such as the current location of the person relative to the robot.

Appendix A

Questionnaires

The appendix lists all questionnaires, as well as consent forms used in the experiments.

A.1 Perception of Robot Behavior

A.1.1 Questionnaire

1. **Name/nickname:** _____
2. **Gender:** □ Male □ Female
3. **Age:** _____
4. **Nationality:** _____
5. **Education:** □ HS □ RS □ Abitur □ FH □ Uni
6. **Do you already have experience in interacting with robots?** □ Yes □ No
7. **If yes, with which robot?** _____

Please rate your impression of the robot on these scales.
Anthropomorphism

8. Fake □—□—□—□—□ Natural
9. Machinelike □—□—□—□—□ Humanlike
10. Unconscious □—□—□—□—□ Conscious
11. Artificial □—□—□—□—□ Lifelike
12. Moving rigidly □—□—□—□—□ Moving elegantly
13. Artificial □—□—□—□—□ Lifelike

Animacy

14. Dead □—□—□—□—□ Alive
15. Stagnant □—□—□—□—□ Lively
16. Mechanical □—□—□—□—□ Organic
17. Artificial □—□—□—□—□ Lifelike
18. Inert □—□—□—□—□ Interactive
19. Apathetic □—□—□—□—□ Responsive

Likeability

20. Dislike □—□—□—□—□ Like

21. Unfriendly □—□—□—□—□ Friendly

22. Unkind □—□—□—□—□ Kind

23. Unpleasant □—□—□—□—□ Pleasant

24. Awful □—□—□—□—□ Nice

Perceived Intelligence

25. Incompetent □—□—□—□—□ Competent

26. Ignorant □—□—□—□—□ Knowledgeable

27. Irresponsible □—□—□—□—□ Responsible

28. Unintelligent □—□—□—□—□ Intelligent

29. Foolish □—□—□—□—□ Sensible

Perceived Safety

30. Anxious □—□—□—□—□ Relaxed

31. Agitated □—□—□—□—□ Calm

32. Quiescent □—□—□—□—□ Surprised

User Satisfaction

33. Frustrating □—□—□—□—□ Exciting

34. Unsatisfying interaction □—□—□—□—□ Satisfying interaction

35. What did NAO do? _____

36. What did NAO want? _____

37. Did NAO want to interact with you? _____

38. Was NAO successful in what it did? _____

39. How did NAO communicate? _____

40. Has NAO recognized or understood something? If yes, what?
_____

41. Is NAO male or female? _____

42. Which are NAO's features that let you recognize its gender?
_____

43. How old is NAO? _____

44. Is NAO too small/big for interacting with it? _____

45. What do you expect NAO can do in the future? _____

46. What would have been different if we would have NAO replaced by a human?
_____

Ich bin damit einverstanden, dass meine Antworten in anonymisiert Form veröffentlicht werden.

A.2 Human Gesture Vocabulary

A.2.1 Consent Form

You are going to participate in an experiment aiming to studying the use of gestures in human-robot interaction, based on the robot waiter scenario. The goal of the experiment is to create a set of gestures people use in their everyday interaction with the waiters in cafes or bars. An action, such as "call the waiter" or "ask for the bill" will be shown on the computer screen. After that, you will need to do a gesture you would typically do in order to achieve that command. Each action is represented with a sentence and words in red represent keywords. You can use one gesture to represent the whole sentence, two gestures, one for each keyword, or a pointing gesture.

A follow-up survey consists of four questions – age, gender, left- or right-handedness and the frequency of your visits to cafes and bars. You will receive one additional point in the final exam of the Automation course for your participation. The experiment typically lasts 10-15 minutes. You are free to stop and leave the experiment at any time, but if you leave before the experiment is concluded you will not receive the additional point.

Your actions will be recorded with a Microsoft Kinect (video and depth camera that provides depth information). Upper body (including face and arms) will be recorded with the video camera and locations of the head, shoulders and hands will be recorded with the depth camera. During the analysis, all the data will be stored in encrypted form on a location inaccessible from the network (both university network and the Internet in general). Your name won't be recorded and your anonymity will be preserved. Recorded images will be used only for the analysis and will not be provided or displayed to third persons. Informed consent forms will be stored separately.

For any further questions you can contact Sasa Bodiroza at bodiroza@bgu.ac.il or bodiroza@informatik.hu-berlin.de.

A.3 Robot Gesture Vocabulary

A.3.1 Consent Form

1. You are going to participate in an experiment aiming to studying the use of gestures in human-robot interaction, based on the robot waiter scenario. The goal of the experiment is to create a set of gestures robots can use in their everyday interaction with the customers in cafes, restaurants or bars. An action, such as "can I suggest you something" or "would you like another one" will be shown on the computer screen. After each action, you will watch 2-3 videos of a small humanoid robot, Aldebaran Nao, performing different gestures associated with the presented action. After that, you will need to rank presented gestures, based on how strongly you associate each gesture with the presented action.

A follow-up survey consists of four questions – age, gender, left- or right-handedness and the frequency of your visits to cafes and bars.

You will receive 1 credit in the class for your participation.

The experiment typically lasts 10-15 minutes. You are free to stop and leave the experiment at any time, but if you leave before the experiment is concluded you will not receive the class credits.

Your name won't be recorded and your anonymity will be preserved.

□ I agree

□ I disagree

A.3.2 Questionnaire

2. **Name/nickname:** _____
3. **Age:** _____
4. **Gender:** □ Male □ Female
5. **How often do you go to cafes or bars?**
   - □ Every day
   - □ 3-4 times per week
   - □ 1-2 times per week
   - □ 1 time every two weeks
   - □ 1 time per month
   - □ Rarely
   - □ Never

Following questions were asked for all gesture alternatives, here labeled with ***n*** for each of the actions, here labeled with ***A***.

Action *A*

6. **Overall impression of the gesture alternative *n*.**
   Very good □—□—□—□—□—□—□—□—□—□ Very bad
7. **Speed of the gesture alternative *n*.**
   Very fast □—□—□—□—□—□—□—□—□—□ Very slow
8. **Precision of the gesture alternative *n*.**
   Very good □—□—□—□—□—□—□—□—□—□ Very bad

In addition, the participants were asked to rank the gesture alternatives of each action (i.e., assigning them values from 1 to 3, to indicate the order from the best to the worst).

Bibliography

Akgün, B., Tunaoğlu, D., and Şahin, E. (2010). Action recognition through an action generation mechanism. In *International Conference on Epigenetic Robotics (EPIROB)*.

Arici, T., Celebi, S., Aydin, A. S., and Temiz, T. T. (2014). Robust gesture recognition using feature pre-processing and weighted dynamic time warping. *Multimedia Tools and Applications*, 72(3):3045–3062.

Baron-Cohen, S. (2001). *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press.

Bartneck, C., Croft, E., and Kulic, D. (2008). Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots. In *Proceedings of the Metrics for Human-Robot Interaction Workshop at the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI 2008)*, pages 37–44. University of Hertfordshire.

Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, 1(1):71–81.

Bazo, D., Vaidyanathan, R., Lentz, A., and Melhuish, C. (2010). Design and testing of a hybrid expressive face for a humanoid robot. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 5317 –5322.

Bergmann, K., Kopp, S., and Eyssel, F. (2010). Individualized gesturing outperforms average gesturing – evaluating gesture production in virtual humans. In Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., and Safonova, A., editors, *Intelligent Virtual Agents*, volume 6356 of *Lecture Notes in Computer Science*, pages 104–117. Springer Berlin Heidelberg.

Berman, S. and Stern, H. (2012). Sensors for gesture recognition systems. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(3):277–290.

Blakemore, S. J., Wolpert, D., and Frith, C. (2000). Why can't you tickle yourself? *Neuroreport*, 11:11–16.

Bodiroža, S., Doisy, G., and Hafner, V. V. (2013a). Position-invariant, real-time gesture recognition based on dynamic time warping. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction*, HRI '13, pages 87–88, Piscataway, NJ, USA. IEEE Press.

Bodiroža, S. and Hafner, V. V. (2014). GRaD: Gesture recognition and disambiguation framework for unconstrained, real-life scenarios. In *Workshop Proceedings of the 13th International Conference on Intelligent Autonomous Systems*, IAS '14, pages 347–353.

Bodiroža, S., Jevtić, A., Lara, B., and Hafner, V. V. (2013b). Learning the relation of motion control and gestures through self-exploration. In *Proceedings of Robotics Challenges and Vision Workshop, at Robotics: Science and Systems*, Berlin, Germany.

Bodiroža, S., Stern, H. I., and Edan, Y. (2012). Dynamic gesture vocabulary design for intuitive human-robot dialog. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '12, pages 111–112, New York, NY, USA. ACM.

Bodiroza, S., Schillaci, G., and Hafner, V. V. (2011). Robot ego-sphere: An approach for saliency detection and attention manipulation in humanoid robots for intuitive interaction. In *Proceedings of the 11th IEEE-RAS Conference on Humanoid Robots*, pages 689–694.

Bolt, R. A. (1980). "put-that-there": Voice and gesture at the graphics interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '80, pages 262–270, New York, NY, USA. ACM.

Burghart, C. R. and Steinfeld, A., editors (2008). *Proceedings of the Metrics for Human-Robot Interaction Workshop at the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI 2008)*.

Burkhard, H., Holzhauer, F., Krause, T., Mellmann, H., Ritter, C., Welter, O., and Xu, Y. (2010). NAO-Team Humboldt 2010. *Humboldt-Universität zu Berlin*.

Celebi, S., Aydin, A. S., Temiz, T. T., and Arici, T. (2013). Gesture recognition using skeleton data with weighted dynamic time warping. In Battiato, S. and Braz, J., editors, *VISAPP 2013 - Proceedings of the International Conference on Computer Vision Theory and Applications, Volume 1, Barcelona, Spain, 21-24 February, 2013.*, pages 620–625. SciTePress.

Chaaraoui, A. A., Padilla-López, J. R., Climent-Pérez, P., and Flórez-Revuelta, F. (2014). Evolutionary joint selection to improve human action recognition with rgb-d devices. *Expert Systems with Applications*, 41(3):786 – 794.

Corradini, A. (2001). Dynamic time warping for off-line recognition of a small gesture vocabulary. In *Proc. IEEE ICCV Workshop on Recognition, Anal., and Tracking of Faces and Gestures in Real-Time Syst.*, RATFG-RTS '01, pages 82 –89.

Dardas, N., Chen, Q., Georganas, N. D., and Petriu, E. (2010). Hand gesture recognition using bag-of-features and multi-class support vector machine. In *Haptic Audio-Visual Environments and Games (HAVE), 2010 IEEE International Symposium on*, pages 1–5.

Dawkins, R. (2013). *The Blind Watchmaker*. Penguin Books Limited.

Dearden, A. and Demiris, Y. (2005). Learning forward models for robots. In *Proceedings of the 19th international joint conference on Artificial intelligence*, IJCAI'05, pages 1440–1445, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Desrochers, S., Morisette, P., and Ricard, M. (1995). Two perspectives on pointing in infancy. In Moore, C. and Dunham, P., editors, *Joint Attention: Its Origins and Role in Development*, pages 85–101. Lawrence Erlbaum Associates.

Doisy, G., Jevtić, A., and Bodiroža, S. (2013). Spatially unconstrained, gesture-based human-robot interaction. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction*, HRI '13, pages 117–118, Piscataway, NJ, USA. IEEE Press.

Doisy, G., Jevtić, A., Lucet, E., and Edan, Y. (2012). Adaptive person-following algorithm based on depth images and mapping. In *Proc. of the IROS Workshop on Robot Motion Planning*.

Ende, T., Haddadin, S., Parusel, S., Wusthoff, T., Hassenzahl, M., and Albu-Schaffer, A. (2011). A human-centered approach to robot gesture based communication within collaborative working processes. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 3367 –3374.

Fleming, K. A., Peters, R. A., and Bodenheimer, R. E. (2006). Image mapping and visual attention on a sensory ego-sphere. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 241–246.

Freeman, W. T. and Roth, M. (1995). Orientation histograms for hand gesture recognition. In *International Workshop on Automatic Face and Gesture Recognition*, volume 12, pages 296–301.

Frintrop, S., Backer, G., and Rome, E. (2005). Selecting what is important: Training visual attention. In *German Conference on Artificial Intelligence*, pages 351–365.

Graham, J. A. and Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology*, 10(1):57–67.

Hafner, V. V. and Schillaci, G. (2011). From field of view to field of reach – could pointing emerge from the development of grasping? *Proceedings of the IEEE Conference on Development and Learning and Epigenetic Robotics (IEEE ICDL-EPIROB 2011), conference abstract in Frontiers in Computational Neuroscience.*

Ham, J., Bokhorst, R., and Cabibihan, J. (2011). The influence of gazing and gestures of a storytelling robot on its persuasive power. In *International Conference on Social Robotics*.

Haruno, M., Wolpert, D. M., and Kawato, M. (1999). Multiple paired forward-inverse models for human motor learning and control. In Michael Kearns, Sara Solla, D. C., editor, *Advances in Neural Information Processing Systems*, volume 11, pages 31–37, Cambridge, MA. MIT Press.

Hegel, F., Gieselmann, S., Peters, A., Holthaus, P., and Wrede, B. (2011). Towards a typology of meaningful signals and cues in social robotics. In *RO-MAN, 2011 IEEE*, pages 72–78, Atlanta, Georgia. IEEE, IEEE.

Ho, C.-C. and MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the godspeed indices. *Computers in Human Behavior*, 26(6):1508 – 1518. Online Interactivity: Role of Technology in Behavior Change.

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge, MA, USA.

Horowitz, D. (1994). Generating rhythms with genetic algorithms. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 2)*, AAAI'94, page 1459, Menlo Park, CA, USA. American Association for Artificial Intelligence.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:194–203.

Iverson, J. M. and Goldin-Meadow, S. (1998). Why people gesture when they speak. *Nature*, 396(6708):228.

Jevtić, A., Doisy, G., Bodiroža, S., Edan, Y., and Hafner, V. V. (2014). Human-robot interaction through 3d vision and force control. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, HRI '14, pages 102–102, New York, NY, USA. ACM.

Jones, J. (2006). Robots at the tipping point: the road to irobot roomba. *Robotics Automation Magazine, IEEE*, 13(1):76 – 78.

Jonsson, G. K. and Thorisson, K. R. (2010). Evaluating multimodal human-robot interaction: A case study of an early humanoid prototype. In *Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research*, MB '10, pages 9:1–9:4. ACM.

Kanda, T., Shiomi, M., Miyashita, Z., Ishiguro, H., and Hagita, N. (2009). An affective guide robot in a shopping mall. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, HRI '09, pages 173–180, New York, NY, USA. ACM.

Kaplan, F. and Hafner, V. V. (2006). The challenges of joint attention. *Interaction Studies*, 7(2):135–169.

Kendon, A. (1986). Current issues in the study of gesture. In Nespoulous, J.-L., Perron, P., and Lecours, A. R., editors, *The Biological Foundations of Gestures: Motor and Semiotic Aspects*, pages 23–47. Lawrence Erlbaum Associates.

Kendon, A. (1988). How gestures can become like words. In Poyatos, F., editor, *Cross-Cultural Perspectives in Nonverbal Communication*, pages 131–141. C. J. Hogrefe.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Gesture: Visible Action as Utterance. Cambridge University Press.

Liebal, K., Pika, S., and Tomasello, M. (2006). Gestural communication of orangutans (pongo pygmaeus). *Gesture*, 6(1):1–38.

Liu, X. and Fujimura, K. (2004). Hand gesture recognition using depth data. In *Proc. IEEE Int. Conf. on Automat. Face and Gesture Recognition (FGR'04)*, pages 529–534.

Mai, N. T. T., Hai, T. T. T., and Son, N. V. (2011). Wizard of Oz for designing hand gesture vocabulary in human-robot interaction. In *Proc. Int. Conf. on Knowledge and Syst. Eng.*, pages 232–238.

Marques, H., Jäntsch, M., Wittmeier, S., Holland, O., Alessandro, C., Diamond, A., Lungarella, M., and Knight, R. (2010). Ecce1: The first of a series of anthropomimetic musculoskeletal upper torsos. In *Humanoid Robots (Humanoids), 2010 10th IEEE-RAS International Conference on*, pages 391–396.

Masataka, N. (2003). From index-finger extension to index-finger pointing: Ontogenesis of pointing in preverbal infants. In Kita, S., editor, *Pointing: Where language, culture, and cognition meet*, pages 69–84. Lawrence Erlbaum Associates, Mahwah, New Jersey, USA.

Mazzei, D., Lazzeri, N., Hanson, D., and De Rossi, D. (2012). Hefes: An hybrid engine for facial expressions synthesis to control human-like androids and avatars. In *Biomedical Robotics and Biomechatronics (BioRob), 2012 4th IEEE RAS EMBS International Conference on*, pages 195 –200.

McNeill, D. (1986). Iconic gestures of children and adults. *Semiotica*, 62(1-2):107–128.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* University of Chicago Press.

McNeill, D., Levy, E. T., and Pedelty, L. L. (1990). Speech and gesture. In Hammond, G. R., editor, *Cerebral control of speech and limb movements*, pages 203–256. Elsevier.

Meltzoff, A. N. and Moore, M. K. (1997). Explaining Facial Imitation: A Theoretical Model. *Early Development and Parenting*, 6(34):179–192.

Mitra, S. and Acharya, T. (2007). Gesture recognition: A survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 37(3):311–324.

Morioka, K., Lee, J.-H., and Hashimoto, H. (2004). Human-following mobile robot in a distributed intelligent sensor network. *IEEE Transactions on Industrial Electronics*, 51(1):229 – 237.

Morris, D. (1979). *Gestures: their origins and distribution.* Cape London.

Nehaniv, C., Dautenhahn, K., Kubacki, J., Haegele, M., Parlitz, C., and Alami, R. (2005). A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction. In *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, pages 371–377.

Ng-Thow-Hing, V., Luo, P., and Okita, S. (2010). Synchronized gesture and speech production for humanoid robots. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 4617 –4624.

Pavlovic, V. I., Sharma, R., and Huang, T. S. (1997). Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:677–695.

Peters, R. A., Hambuchen, K. E., Kawamura, K., and Wilkes, D. M. (2001). The sensory ego-sphere as a short-term memory for humanoids. In *Proceedings of the IEEE-RAS Conference on Humanoid Robots*, pages 451–460.

Pika, S., Liebal, K., and Tomasello, M. (2003). Gestural communication in young gorillas (gorilla gorilla): Gestural repertoire, learning, and use. *American Journal of Primatology*, 60(3):95–111.

Poggi, I. (2002). From a typology of gestures to a procedure for gesture production. In *Revised Papers from the International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction*, GW '01, pages 158–168, London, UK, UK. Springer-Verlag.

Posner, M. I., Rafal, R. D., Choate, L. S., and Vaughan, J. (1985). Inhibition of return: Neural basis and function. *Cognitive Neuropsychology*, 2(3):211–228.

Potkonjak, V., Svetozarevic, B., Jovanovic, K., and Holland, O. (2011). Anthropomimetic robot with passive compliance - contact dynamics and control. In *Control Automation (MED), 2011 19th Mediterranean Conference on*, pages 1059 –1064.

Quek, F. K. (1995). Eyes in the interface. *Image and Vision Computing*, 13(6):511 – 525.

Rautaray, S. and Agrawal, A. (2012). Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, pages 1–54.

Rea, D., Young, J., and Irani, P. (2012). The roomba mood ring: An ambient-display robot. In *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*, pages 217 –218.

Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J., and Pfeifer, R. (2008). Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub. In *IEEE International Conference on Robotics and Automation, 2008. ICRA 2008*, pages 962–967.

Sadeghipour, A. and Kopp, S. (2009). A probabilistic model of motor resonance for embodied gesture perception. In Ruttkay, Z., Kipp, M., Nijholt, A., and Vilhjálmsson, H. H., editors, *Intelligent Virtual Agents*, volume 5773 of *Lecture Notes in Computer Science*, pages 90–103. Springer Berlin Heidelberg.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49.

Salem, M., Kopp, S., Wachsmuth, I., and Joublin, F. (2010). Towards an integrated model of speech and gesture production for multi-modal robot behavior. In *RO-MAN, 2010 IEEE*, pages 614 –619.

Schillaci, G., Bodiroža, S., and Hafner, V. V. (2013). Evaluating the effect of saliency detection and attention manipulation in human-robot interaction. *International Journal of Social Robotics*, 5(1):139–152.

Schillaci, G. and Hafner, V. V. (2011a). Prerequisites for intuitive interaction – on the example of humanoid motor babbling. In *Proceedings of the Workshop on the Role of Expectations in Intuitive Human-Robot Interaction (HRI 2011)*, pages 23–27.

Schillaci, G. and Hafner, V. V. (2011b). Random movement strategies in self-exploration for a humanoid robot. In *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2011)*, pages 245–246.

Schillaci, G., Lara, B., and Hafner, V. (2012). Internal simulations for behaviour selection and recognition. In Salah, A., Ruiz-del Solar, J., Meriçli, c., and Oudeyer, P.-Y., editors, *Human Behavior Understanding*, volume 7559 of *Lecture Notes in Computer Science*, pages 148–160. Springer Berlin Heidelberg.

Schlömer, T., Poppinga, B., Henze, N., and Boll, S. (2008). Gesture recognition with a wii controller. In *Proceedings of the 2Nd International Conference on Tangible and Embedded Interaction*, TEI '08, pages 11–14, New York, NY, USA. ACM.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., and Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology*, 6(4):147–151.

Schulz, R., Stockwell, P., Wakabayashi, M., and Wiles, J. (2006). Towards a spatial language for mobile robots. In *Proceedings of the 6th International Conference on the Evolution of Language*, pages 291–298.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognition*, pages 1297–1304, Colorado Springs, CO, USA.

Smith, J. R. (1991). Designing biomorphs with an interactive genetic algorithm. *International Conference on Genetic Algorithms*, pages 535–538.

Steels, L. (1998). The origins of syntax in visually grounded robotic agents. *Artificial Intelligence*, 103(1–2):133 – 156. <ce:title>Artificial Intelligence 40 years later</ce:title>.

Steels, L. and Spranger, M. (2012). Emergent mirror systems for body language. In Steels, L., editor, *Experiments in Cultural Language Evolution*, pages 87 – 109. John Benjamins.

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., and Goodrich, M. (2006). Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, HRI '06, pages 33–40. ACM.

Stern, H., Wachs, J., and Edan, Y. (2006). Human factors for design of hand gesture human - machine interaction. In *Systems, Man and Cybernetics, 2006. SMC '06. IEEE International Conference on*, volume 5, pages 4052–4056.

Stern, H. I., Wachs, J. P., and Edan, Y. (2008). Designing hand gesture vocabularies for natural interaction by combining psycho-physiological and recognition factors. *International Journal of Semantic Computing*, 02(01):137–160.

Suarez, J. and Murphy, R. (2012). Hand gesture recognition with depth images: A review. In *RO-MAN, 2012 IEEE*, pages 411–417.

Takayama, L. and Pantofaru, C. (2009). Influences on proxemic behaviors in human-robot interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5495 –5502.

Takemura, N. and Inui, T. (2011). A developmental model of infant reaching movement: Acquisition of internal visuomotor transformations. In Wang, R. and Gu, F., editors, *Advances in Cognitive Neurodynamics (II)*, pages 135–138. Springer Netherlands.

ten Holt, G. A., Reinders, M. J. T., and Hendriks, E. A. (2007). Multi-dimensional dynamic time warping for gesture recognition. In *Proc. of the Conf. of the Advanced School for Computing and Imaging*.

Thrun, S., Beetz, M., Bennewitz, M., Burgard, W., Cremers, A., Dellaert, F., Fox, D., Haehnel, D., Rosenberg, C., Roy, N., et al. (2000). Probabilistic algorithms and the interactive museum tour-guide robot minerva. *The International Journal of Robotics Research*, 19(11):972–999.

Tomasello, M. (1995). Joint attention as social cognition. In Moore, C. and Dunham, P. J., editors, *Joint Attention: Its Origins and Role in Development*, pages 103–130. Lawrence Erlbaum Associates.

Tomasello, M., Call, J., Warren, J., Frost, G. T., Carpenter, M., and Nagell, K. (1997). The ontogeny of chimpanzee gestural signals: A comparison across groups and generations. *Evolution of Communication*, 1(2):223–259.

Treisman, A. (1985). Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing*, 31(2):156–177.

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97 – 136.

Vallejo, G., Fernández, M. P., Tuero, E., and Livacic-Rojas, P. (2010). Análisis de medidas repetidas usando métodos de remuestreo (analyzing repeated measures using resampling methods). *Anales de Psicologia*, 26(2):400–409.

van Oosterhout, T. and Visser, A. (2008). A visual method for robot proxemics measurements. *Proceedings of the Metrics for Human-Robot Interaction Workshop in affiliation with the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI 2008), Technical Report 471*, pages 61–68.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *Int. J. Comput. Vision*, 57:137–154.

Wachs, J. P., Kölsch, M., Stern, H., and Edan, Y. (2011). Vision-based hand-gesture applications. *Commun. ACM*, 54(2):60–71.

Wachsmuth, I. and Kopp, S. (2002). Lifelike gesture synthesis and timing for conversational agents. In *Revised Papers from the International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction*, GW '01, pages 120–133, London, UK, UK. Springer-Verlag.

Walker, M. B. and Nazmi, M. K. (1979). Communicating shapes by words and gestures. *Australian Journal of Psychology*, 31(2):137–143.

Wolpert, D. and Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, 11(7-8):1317–1329.

Yamato, J., Ohya, J., and Ishii, K. (1992). Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on*, pages 379–385.

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Berlin, den February 14, 2017 .....

Statement of authorship

I declare that I completed this thesis on my own and that information which has been directly or indirectly taken from other sources has been noted as such.

Berlin, February 14, 2017 .....